Data-Driven Innovation

BIG DATA FOR GROWTH AND WELL-BEING



OECD

Data-Driven Innovation

BIG DATA FOR GROWTH AND WELL-BEING



BETTER POLICIES FOR BETTER LIVES

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Please cite this publication as:
OECD (2015), *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.
http://dx.doi.org/10.1787/9789264229358-en

ISBN 978-92-64-22934-1 (print)
ISBN 978-92-64-22935-8 (PDF)

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Photo credits: Cover © montage by Baseline Arts

Corrigenda to OECD publications may be found on line at: *www.oecd.org/about/publishing/corrigenda.htm*.

© OECD 2015

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgement of OECD as source and copyright owner is given. All requests for public or commercial use and translation rights should be submitted to *rights@oecd.org*. Requests for permission to photocopy portions of this material for public or commercial use shall be addressed directly to the Copyright Clearance Center (CCC) at *info@copyright.com* or the Centre français d'exploitation du droit de copie (CFC) at *contact@cfcopies.com*.

Preface

Social and economic activities are increasingly migrating to the Internet. The cost of data collection, storage and processing continues to decline dramatically. Ever larger volumes of data will be generated from the Internet of Things, smart devices, and autonomous machine-to-machine communications. We are now at the cusp of a new era, in which "big data" will play a transformative role.

The "datafication" of the economy and society holds many promises in a wide range of areas, from health to agriculture, from public governance to innovation, and from education to the environment, to name just a few. The "low-hanging fruit" of data-driven innovation (DDI) may be clear, but the full scope of potential benefits is much more difficult to grasp, resulting in opportunities that may be lost.

Seizing these benefits poses a formidable challenge to policymakers. In the years ahead, the pivot to a data-driven world will have important implications for policies ranging from privacy, consumer policy, competition, taxation, innovation and especially jobs and skills.

We will need, for example, to recast how we think about infrastructure in the 21st Century, and expand it to encompass broadband networks, cloud computing and data itself. Ensuring that DDI leads to growth will require focusing on small and medium enterprises and high value-added services, such as design and engineering. The questions of access and ownership of data are also essential. Governments will need to understand and strike the right balance between the social benefits of "openness", and individuals' and organisations' legitimate concerns about such openness.

As well as a catalyst for growth, innovation and productivity gains, DDI will be a disruptive force, with far-reaching effects on the economy and well-being. Policymakers will need to consider the trade-offs, complementarities and possible unintended consequences both of their policy actions - and of inaction. We need to ensure that the benefits of DDI are widely shared, and that far from creating new divides they do not leave anyone behind.

This will be no easy task. This report helps policymakers to be proactive, instead of reactive, by outlining these trade-offs. It uses the breadth of the OECD's expertise to outline the contours of this phenomenon and frames a number of the policy dialogues that need to occur so as to fully benefit from the coming era of ubiquitous data.



Angel Gurría
Secretary-General
OECD

Foreword

Early in 2011 the OECD began a project on *New Sources of Growth: Knowledge-based Capital* (KBC). The project was inspired by findings from the OECD's *Innovation Strategy*, originally published in 2010 and now updated to 2015 (forthcoming). According to these findings, many innovating firms invest, beyond R&D, in a broader range of intangibles assets including i) intellectual property (e.g. patents, trademarks, copyrights, trade secrets, designs); ii) digital data and information (e.g. data and analytics); and iii) economic competencies (e.g. organisational capital and firm-specific skills). These intangible assets are referred to as knowledge-based capital (KBC).

This report focuses on digital data and analytics and their effects on innovation, growth and well-being. It aims to improve the evidence base on the role of data-driven innovation (DDI) in boosting productivity growth and contributing to well-being. It also offers policy guidance for maximizing the benefits of DDI and mitigating the associated economic and societal risks. The insights in the report are intended to help policy makers better understand DDI, incorporate its multidimensionality into policy design and "identify trade-offs, complementarities and unintended consequences of policy choices". This report contributes to the goal of building and maintaining "resilient economies and inclusive societies" while enhancing the productivity and competitiveness of industries, as articulated in the OECD Ministerial Council Statements of 2014 and 2015.

The work on DDI has drawn on expertise from different directorates within the OECD. Supported with financial resources from the Secretary-General's Central Priority Fund and in-kind contributions from the Netherlands, the Directorate for Science, Technology and Innovation led the two-year effort. Other partners have been the Directorate for Employment, Labour and Social Affairs, and the Directorate for Public Governance and Territorial Development. Owing to this co-operative effort, the publication's different chapters were discussed and declassified by various OECD committees, including the Committee on Digital Economy Policy which had oversight responsibility for the project; the Committee on Consumer Policy; the Committee for Scientific and Technological Policy; the Health Committee; and the Public Governance Committee. The comments and inputs received from delegates to these official OECD bodies are gratefully acknowledged.

The material presented in this book will feed ongoing and future OECD projects, most notably the OECD project on the Next Production Revolution (NPR, http://oe.cd/npr). Further information on the work on DDI, including follow-up work, will be available on the OECD website, at http://oe.cd/bigdata.

ACKNOWLEDGEMENTS

The work on Data-Driven Innovation represents an OECD collective effort led and co-ordinated by Christian Reimsbach-Kounatze (Information Economist and Policy Analyst, Division for Digital Economy Policy) under the guidance and oversight of Andrew Wyckoff (Director, Directorate for Science, Technology and Innovation). Jørgen Abild Andersen (Denmark), Chair of the OECD Committee on Digital Economy Policy, and Anne Carblanc (Head of Division, Division for Digital Economy Policy) provided directions and advice throughout the process.

Chapters 1, 2, 3, 4 and 6 were authored by Mr. Reimsbach-Kounatze. In particular, Chapter 1 ("The phenomenon of data-driven innovation") benefited from input from Sabine Gerdon; Chapter 2 ("Mapping the global data ecosystem and its points of control") from Andrea de Panizza and the Netherlands Organisation for Applied Scientific Research (TNO – Jop Esmeijer, Bas Kotterink, Anne F. van Veenstra, Tom Bakker, Merel Ooms, Anna van Nunen, and Silvain de Munck); Chapter 3 ("How data now drive innovation") from Rudolf van der Berg; and Chapter 6 ("Skills and employment in a data-driven economy") from Cristina Serra Vallejo, Sabine Gerdon and the Research Institute for Applied Knowledge Processing, Germany (FAW/n – Estelle L.A. Herlyn, Thomas Kämpke, Franz J. Radermacher, and Dirk Solte). Chapter 5 ("Building trust for data-driven innovation") was authored by Laurent Bernat and Michael Donohue. Chapter 7 ("Promoting data-driven scientific research") was written by Giulia Ajmone Marsan, with guidance from Mario Cervantes. Chapter 8 ("The evolution of health care in a data-rich environment") was authored by Jillian Oderkirk and Elettra Ronchi. Chapter 9 ("Cities as hubs for data-driven innovation") was written by David Gierten, with input from TNO. Chapter 10 ("Governments leading by example with public sector data") was authored by Barbara Ubaldi, with contributions from Graham Vickery. Randall Holden edited the book and Janine Treves, James Arkinstall and Kate Brooks provided support with the overall presentation.

Some chapters benefited from additional expertise within the OECD through extensive rounds of comments. Special thanks therefore go to: John Davies (Competition Division of the Directorate for Financial and Enterprise Affairs); Vincenzo Spiezia (Economic Analysis and Statistic Division of the Directorate for Science, Technology and Innovation); Jesse Eggert, Eric Robert and Liz Chien (Digital Economy Team of the Center for Tax Policy and Administration); and Guillermo Montt (Division for Employment Analysis and Policy of the Directorate for Employment, Labour and Social Affairs).

Analysis and policy conclusions also benefited from advice provided by an international panel of independent experts including Devdatt Dubhashi (Professor, Department of Computer Science and Engineering, Chalmers University of Technology), Brett Frischmann (Director, Cardozo Intellectual Property & Information Law Program and Professor of Law, Benjamin N. Cardozo School of Law), Jakob Haesler (Co-founder, tinyclues), Simon Hania (Corporate Privacy Officer, TomTom), and Sarah Spiekermann (Head of the Institute for Management Information Systems,

Vienna University of Economics and Business). Thanks also go to Brian Kahin (Fellow at the MIT Sloan School Center for Digital Business) for his very informative comments. In addition, the report benefited from the advice of a panel of delegates drawn from the participating OECD committees. Many thanks go to Andre Loth (France), Emilio Garcia Garcia and Ruth Del Campo Becares (Spain), Tony O'Connor (United Kingdom), Hugh Stevenson and Stacy Feuer (United States), and Robin Wilton (Internet Technical Advisory Committee to the OECD).

The work on DDI and this book also benefited from discussions with the authors of some of the most prominent literature on "big data". Thanks therefore go to: Kenneth Cukkier and Viktor Mayer-Schönberger, the authors of "Big Data: A Revolution That Will Transform How We Live, Work and Think"; Robert Kirkpatrick and his team at United Nations Global Pulse for their work on big data for well-being and development; Carl Kalapesi and Joel Nicholson at the World Economic Forum (WEF) for their work on personal data and big data for development; Hasan Bakhshi and Juan Mateos-Garcia at Nesta for their work on the "datavores"; and Paul Hofheinz at the Lisbon Council and Michael Mandel at the Progressive Policy Institute (PPI) for their work on the transatlantic policy issues raised by big data.

Finally, two major events helped scope, develop and test analytic and policy ideas with academics, policymakers and practitioners. One was the *2012 Technology Foresight Forum* (http://oe.cd/tff2012), held at the OECD headquarters in Paris, France, on 22 October 2012; The second event was the *4th Global Forum on the Knowledge Economy* (GFKE, http://oe.cd/gfke2014), held in Tokyo, Japan, on 2-3 October 2014, and co-organised with, and hosted by, the Japanese Ministry for Internal Affairs and Communications and the Japanese Ministry of Economy, Trade and Industry. Special thanks go to the hosts, to Hajime Oiso, Aki Irie and Yuka Miyazaki who helped organise the GFKE, and to all the participants of both events.

Table of contents

Chapter 1. The phenomenon of data-driven innovation 19

1.1. The rise of "big data" and data-driven innovation 22
1.2. Objectives and structure of this volume 33
1.3. Common key challenges and policy considerations 53
Annex – Highlights of the 2014 Global Forum on the Knowledge Economy 55

Chapter 2. Mapping the global data ecosystem and its points of control 69

2.1. The key actors and their main technologies, services and business models 71
2.2. Interactions in the data ecosystem 91
2.3. Key challenges in the global data ecosystem 98
2.4. Key findings and policy conclusions 112
Annex – OECD (1985) Declaration on Transborder Data Flows 115

Chapter 3. How data now drive innovation 131

3.1. The exponential growth in data generated and collected 133
3.2. The pervasive power of data analytics 143
3.3. From informing to driving decision-making 150
3.4. Key findings and policy conclusions 159

Chapter 4. Drawing value from data as an infrastructure 177

4.1. Data as infrastructural resource 179
4.2. The economics of data 184
4.3. Towards a data governance framework for better data access, sharing and interoperability 186
4.4. Key findings and policy conclusions 197

Chapter 5. Building trust for data-driven innovation 207

5.1. Security for data-driven innovation 208
5.2. Privacy protection for data-driven innovation 216
5.3. Key findings and policy conclusions 227

Chapter 6. Skills and employment in a data-driven economy 237

6.1. "Creative destruction" in labour markets 239
6.2. The growing importance of data specialist skills and employment 252
6.3. Promoting data-driven innovation and smoothing structural change 272
6.4. Key findings and policy conclusions 280
Annex – Selected statistical definitions of data specialist occupations 282

Chapter 7. Promoting data-driven scientific research 299

7.1. The evolving scientific enterprise 301
7.2. Impacts of open access to science, research and innovation 306
7.3. Policies and practices: OECD countries and beyond 315
7.4. Key findings and policy conclusions 322

Chapter 8. The evolution of health care in a data-rich environment 331

8.1. Drivers of growth of digitised health data 333
8.2. Data-driven innovation to improve health care quality and health system performance 337
8.3. Data-driven innovation for smarter models of care 344
8.4. Transforming health research with big data 348
8.5. Critical success factors and policy priorities 355
8.6. Key findings and policy conclusions 363

Chapter 9. Cities as hubs for data-driven innovation 379

9.1. The urban data ecosystem 380
9.2. Opportunities for data-driven innovation in cities 382
9.3. Policy priorities 389
9.4. Key findings and policy conclusions 395

Chapter 10. Governments leading by example with public sector data 403

10.1. The potential of public sector data 406
10.2. Key challenges in implementing open data strategies 416
10.3. Key findings and policy conclusions 434
Annex – Principles of the OECD (2008) Council Recommendation on PSI 437

Glossary 449

Figures

Figure 1.1. Estimated worldwide data storage 20
Figure 1.2. Investment in physical and knowledge-based capital, 2010 22
Figure 1.3. Average revenue per employee of top 250 ICT firms, 2000-13 23
Figure 1.4. Big data-related financing activities, Q1 2008-Q4 2012 24
Figure 1.5. Top locations by number of co-location data centres and top sites hosted 25
Figure 1.6. Trends in data intensity of the Canadian and United States economies, 1999-2013 27
Figure 1.7. The data value cycle 33
Figure 1.8. The diffusion of selected ICT tools and activities in enterprises, 2013 37
Figure 1.9. Data analytics related articles in the Science Direct repository, 1995-2014 43
Figure 2.1. Main phases of the data value cycle with their key types of actors 71
Figure 2.2. The data ecosystem as layers of key roles of actors 72
Figure 2.3. Market prices per record for personal data by type, 2011 83
Figure 2.4. Partnerships in the Hadoop ecosystem, January 2013 92
Figure 2.5. OECD and major exporters of ICT services, 2000 and 2013 98
Figure 2.6. App switching costs by platform and by country, 2012 106
Figure 3.1. DDI: The data value cycle and confluence of key trends and enabling factors 132
Figure 3.2. The diffusion of online purchases, 2013 and 2007 134
Figure 3.3. Monthly global Internet Protocol (IP) data traffic, 2005-17 135
Figure 3.4. OECD wireless broadband penetration, by technology, December 2009 and June 2013 136
Figure 3.5. Local content sites hosted in country, 2013 137
Figure 3.6. The diffusion of RFID in enterprises, 2011 140
Figure 3.7. Patents on M2M, data analytics and 3D printing technologies, 2004-14 141
Figure 3.8. Machine-to-machine applications and technologies, by dispersion and mobility 142

Figure 3.9. Average data storage cost for consumers, 1998-2012 145
Figure 3.10. Cost of genome sequencing, 2001-14 146
Figure 3.11. Enterprises using cloud computing services by *employment size class*, 2014 149
Figure 3.12. Algorithmic trading as a share of total trading 156
Figure 3.13. Fever estimations in the United States, January 2011-December 2012 159
Figure 4.1. The data common continuum 191
Figure 5.1. Digital security risk management cycle 215
Figure 6.1. Labour productivity and employment in selected OECD countries (1950-2011) 242
Figure 6.2. Trends in the share of ICT specialists in selected OECD countries, 2003-13 243
Figure 6.3. Index of changing work tasks in the United States 250
Figure 6.4. Firms using innovation-relevant skills, 2008-10 253
Figure 6.5. Main phases of the data value cycle with their key types of data specialist occupations 255
Figure 6.6. Data and ICT specialists in context 256
Figure 6.7. Data specialists in selected OECD countries, 2011-13 257
Figure 6.8. Trends in the share of data specialists in the United States, 1999-2013 258
Figure 6.9. Trends in the share of data specialists in total employment in Canada, 1999-2014 258
Figure 6.10. Trends in relative average wage of data specialists in the United States, 1999-2013 259
Figure 6.11. Trends in relative average wage of data specialists in Canada, 1998/99-2013/14 259
Figure 6.12. Data specialist jobs outlook in the United States, 2012-22 260
Figure 6.13. Distribution of data specialists per industry in selected OECD countries, 2013 261
Figure 6.14. Data-related tertiary graduates, by gender, 2005 and 2012 262
Figure 6.15. Growth of job starters listed in LinkedIn with a focus on data analytics and data science 267
Figure 6.16. Data specialist skills and competence mix 270
Figure 6.17. Trends in the number of certified/professional privacy and security experts, 2003-13 270
Figure 6.18. Level of proficiency in problem solving in technology-rich environments, 2012 275
Figure 6.19. Science, reading and mathematics proficiency at age 15, 2009 276
Figure 6.20. STEM (science, technology, engineering and mathematics) graduates 277
Figure 6.21. STEM graduates by disciplines, 2012 278
Figure 8.1. Planned and implemented uses of data from electronic health record systems 339
Figure 8.2. Smart mobile health (mHealth) applications 347
Figure 8.3. Risks associated with the collection and use of personal health data 356
Figure 9.1. Urban data categories 380
Figure 9.2. Key actors handling proprietary and open data in cities 393
Figure 10.1. The relationship between public sector information and open government data 405
Figure 10.2. Variety of data sets in the centralised government portal 406
Figure 10.3. Main objectives of open government data strategies 407
Figure 10.4. Open government data's main challenges as reported by countries 416

Tables

Table 2.1. Performance of the top Internet firms involved in the Hadoop ecosystem, 2013 93
Table 2.2. Performance of the top ICT service and software firms involved in the Hadoop ecosystem, 2013 93
Table 2.3. Performance of the top ICT hardware firms involved in the Hadoop ecosystem, 2013 94
Table 6.A1 Europe: Occupations included in the operational definition of the Data specialists 282
Table 6.A2 United States: Occupations included in the operational definition of data specialist 282
Table 6.A3 Australia: Occupations included in the operational definition of data specialist 282
Table 6. A4 Canada: Occupations included in the operational definition of data specialist 282
Table 8.1. Number of countries reporting data and data linkages 338
Table 9.1. Life cycles of selected technologies, networks and infrastructures 390
Table 10.1. Machine-readability, open formats and interoperability 419
Table 10.2. Budgeting for the costs of opening up public sector information 422
Table 10.3. Public sector information licensing practices 430

Abbreviations

AD	Alzheimer's disease
ADRN	Administrative Data Research Centres, United Kingdom
AIC triad	Availability, integrity and/or confidentiality of information
AMI	Acute myocardial infarction
APIs	Application programming interfaces
ATS	Algorithmic trading systems
BiOS Initiative	Biological Innovation for Open Society
BPP	Billion Price Project
CAGR	Compound annual growth rate
CancerLinQ	American Society of Clinical Oncology's Cancer Learning Intelligence Network for Quality
CCD	Ciudad Creativa Digital, Guadalajara, Mexico
CCLA	City Climate Leadership Award
ccTLDs	Country code top-level domains
CDC	Centers for Disease Control and Prevention, United States
CDNs	Content delivery networks
CEPS	Centre for European Policy Studies
CER	Comparative effectiveness research
CLA	Contributor Licence Agreement
CODATA	Committee on Data for Science and Technology
CONIYT	National Commission of Technological Research, Chile
CSV	Comma-separated values
CT	Computed tomography
DBMS	Database management system
DDI	Data-driven innovation
DEC	Department of Environmental Conservation, New York State
DoS	Denial of service
DRM	Digital rights management
DSSs	Decision support systems
DW	Data warehouse
EBI	European Bioinformatics Institute
ECHO	European Collaboration for Healthcare Optimization
EDF	Électricité de France
EDW	Enterprise data warehouse
EHRs	Electronic health records
EITC	Earned income tax credits
EMBL	European Molecular Biology Laboratory
EMIF	European Medical Information Framework
ENoLL	European Network of Living Labs
EP	European Parliament
EPRs	Electronic personal records
ERDF	Électricité Réseau Distribution France

ERP	Enterprise resource planning
Esri	Environmental Systems Research Institute
ESSC	European Statistical System Committee
ETDE	Energy Technology Data Exchange
eTRIKS	Delivering European Translational Information & Knowledge Management Services
EU-ADR	EU Advanced Drug Reporting initiative or should ADR really be "Adverse Drug Reactions"? Please confirm
EUNOIA	Evolutive User-centric Networks for Intraurban Accessibility, European Union
EuroHOPE	European Health Care Outcomes, Performance and Efficiency Project
Fing	Fondation Internet Nouvelle Génération, France
fMRI	Functional magnetic resonance imaging
GfK	Gesellschaft für Konsumforschung, Society for Consumer Research
GIS	Geographic information systems
GP	General practitioner
GPHIN	Global Public Health Information Network
GPS	Global positioning system
HCQI	Health Care Quality Indicators (OECD)
HDDs	Hard disk drives
HES	Hospital Episode Statistics
HGF	Hypothesis generation framework
HMO	Health care maintenance organisation
IaaS	Infrastructure as a service
IAPP	International Association of Privacy Professionals
ICC	Integrated circuit card
ICES	Institute for Clinical and Evaluative Sciences, Canada
ICGC	International Cancer Genome Consortium
ICS-CERT	Industrial Control System Cyber Emergency Response Team, United States
ICSTI	International Council for Scientific and Technical Information
ICSU	International Council for Science
ICTs	Information and communication technologies
IGOs	International governmental organisations
IEA	International Energy Agency
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
INTEGRATE	Integrative Cancer Research through Innovative Biomedical Infrastructures, European Commission
IoT	Internet of Things
IPRs	Intellectual property rights
ISO	International Organization for Standardization
ISPs	Internet service providers
ITU	International Telecommunication Union
JSON	JavaScript Object Notation
Kbit	Kilobit, equals 1 000 bits
M2M	Machine-to-machine (communication)

Mbit	Megabit, equals 1 000 000 bits
MGI	McKinsey Global Institute
MOOCs	Massive open online courses
MR	Magnetic resonance
NCDs	Non-communicable diseases
NDES	National Digital Economy Strategy
NFC	Near field communication
NHGRI	National Human Genome Research Institute
NIT	Negative income tax
NLP	Natural language processing
NPISHs	Non-profit institutions serving households
NSF	National Science Foundation, United States
OCR	Optical character recognition
ODbL	Open Database License
ODI	Open Data Institute, United Kingdom
OLAP	Online analytical processing
OLTP	Online transaction processing
OSS	Open source software
OSTP	Office of Science and Technology Policy, United States
PaaS	Platform as a service
PAW Conference	Predictive Analytics World Conference
PB	Petabytes
PCOR	Patient-centred health outcomes research
PCT	Patent Cooperation Treaty
PCTs	Primary care trusts
PEDW	Patient Episode Database for Wales
PET	Positron emission tomography
PGETIC	Global Strategic Plan for Rationalisation of ICT Costs in Public Administration, Portugal
PII	Personal identifying information
PNAS	Proceedings of the National Academy of Sciences
ProMED	Program for Monitoring Emerging Diseases
PSI	Public sector information
QIN	Quantitative Imaging Network, United States
QoS	Quality of service
RDA	Research Data Alliance
RDF	Resource Description Framework
RFID	Radio frequency identification
RPAS	Remote piloted aircraft systems
SaaS	Software as a service
SALUS	Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies
SCOAP3	Sponsoring Consortium for Open Access Publishing in Particle Physics
Sense-OS	Sense Observation Systems
SGDR	Sui generis database right
SHAs	Strategic health authorities
SIM	Subscriber identity module
SIS	Swedish Standardization Institute
SNA	UN System of National Accounts

SPARC	Scholarly Publishing and Academic Resources Coalition
SPECT	Single photon emission computed tomography
SQL	Structured query language
SSDs	Solid-state drives
STRIDE	Stanford Translational Research Integrated Database Environment
TCGA	The Cancer Genome Atlas
TfL	Transport for London
TNO	Netherlands Organisation for Applied Scientific Research
TRANSFoRm	Translational Research and Patient Safety in Europe
TRIPS	Trade-Related Aspects of Intellectual Property Rights
UAV	Unmanned aerial vehicles
UKDBIS	UK Department for Business Innovation & Skills
URIs	Uniform resource identifiers
USPTO	United States Patent and Trademark Office
VRM	Vendor Relationship Management
WCT	WIPO Copyright Treaty
WIPO	World Intellectual Property Organization
WITSA	World Information Technology and Services Alliance
WPA	Wireless personal area
WT	Wellcome Trust
XML	eXtensible Markup Language

Executive summary

Close to real-time analysis of large volumes of data (big data) – generated from a myriad of transactions, production and communication processes – is accelerating knowledge and value creation across society to unforeseen levels. Data-driven innovation (DDI) refers to significant improvement of existing, or the development of new, products, processes, organisational methods and markets emerging from this phenomenon.

DDI has the potential to enhance resource efficiency and productivity, economic competitiveness, and social well-being as it begins to transform all sectors in the economy, including low-tech industries and manufacturing. The exploitation of DDI has already created significant value-added for many businesses and individuals, and more can be expected to follow. Some estimates put the global market for big data related technology and services at USD 17 billion in 2015, with a growth rate of 40% on average every year since 2010. Available evidence also shows that firms using DDI have raised productivity faster than non-users by around 5-10%.

DDI can also help address social and global challenges, including climate change and natural disasters, health and ageing populations, water, food, energy security, and mass urbanisation. Investments in public administration, research and education, and health care will be particularly fruitful in the short term, as these areas rely heavily on the collection and analysis of information, but still face a relatively low level of computerisation in most countries.

The disruptive nature of DDI requires addressing major economic and societal challenges and calls for a whole-of-government and participatory approach to help maximise the benefits and mitigate associated risks and obstacles.

Two clusters of challenges should be met by policy makers in the transition towards a data-driven economy:

1. Governments should consider addressing the negative effects of "creative destruction" while stimulating investments in:
   - *the infrastructure needed for DDI,* particularly in mobile broadband, cloud computing, the Internet of Things, and data, with a strong focus on small and medium-sized enterprises (SMEs) and high value-added services
   - *the public sector, health care, science and education* to pick the "low-hanging fruit" that can boost efficiency, knowledge sharing and well-being in the short term, and help better address global challenges
   - *organisational change and entrepreneurship in the private and public sector* by encouraging a culture of data-driven experimentation and learning
   - *continuous education training and skills development beyond science, technology, engineering and mathematics (STEM) fields* to take advantage of job creation opportunities and smooth structural change while addressing inequality in earnings in labour markets.

2. Governments should aim to understand and strike the right balance between the social benefits of "openness" and individuals' and organisations' legitimate concerns of such openness by encouraging:

   - *the free flow of data across nations and organisations*. This also includes ensuring that the Internet remains an open platform for innovation; promoting both open access to data and interoperability of data-driven services; and empowering actors to reuse their data across interoperable applications (i.e. data portability).
   - *the responsible usage of personal data and the prevention of harm caused by privacy violations.* This also includes enhancing the participation of individuals; the transparency of data processing; the effectiveness of privacy enforcement; and the adoption of a privacy risk management approach.
   - *a culture of digital risk management across society*, involving all stakeholders of the data ecosystem.
   - *data sharing and the appropriation of returns on investments (ROI)* through a combination of alternative incentive mechanisms such as data citations and intellectual property rights (IPR) licences that enable sharing such as Creative Commons and open source software licences.
   - *coherent assessment of market concentration and competition barriers* through better definitions of the relevant market and the consideration of potential consumer detriments due to privacy violation. This will also require a better dialogue between regulatory authorities (in particular in the area of competition, privacy and consumer protection).
   - *improved measurement* to help better assess the economic value of data assets, prevent base erosion and profit shifting (BEPS), and design better DDI policies.

In addressing these two clusters of challenges, policy makers should acknowledge that DDI may favour concentration and greater information asymmetry and with that, shifts in power: away from individuals to organisations; from traditional businesses to data-driven businesses; and from governments to data-driven businesses (the latter can gain more knowledge about citizens than governments). These shifts could exacerbate existing inequalities and lead to a new digital (data) divide that could undermine social cohesion and economic resilience if not addressed.

Given all of this, governments have an important role to play in promoting DDI and mitigating the associated risks.

Chapter 1

The phenomenon of data-driven innovation

This chapter provides a synthesis of the main findings of Phase II of the OECD project on New Sources of Growth: Knowledge-Based Capital*, in particular its pillar which focuses on data-driven innovation (KBC2: DATA). It first presents available evidence on the increasing role of "big data" and data analytics, highlighting in particular the potential of data-driven innovation (DDI) for economic growth, development, and well-being. It then presents the context and policy issues related to the various aspects of DDI covered in this book, chapter by chapter. The discussion concludes by raising key challenges that most countries will face as DDI takes off and accelerates, and the policy considerations they will need to address.*

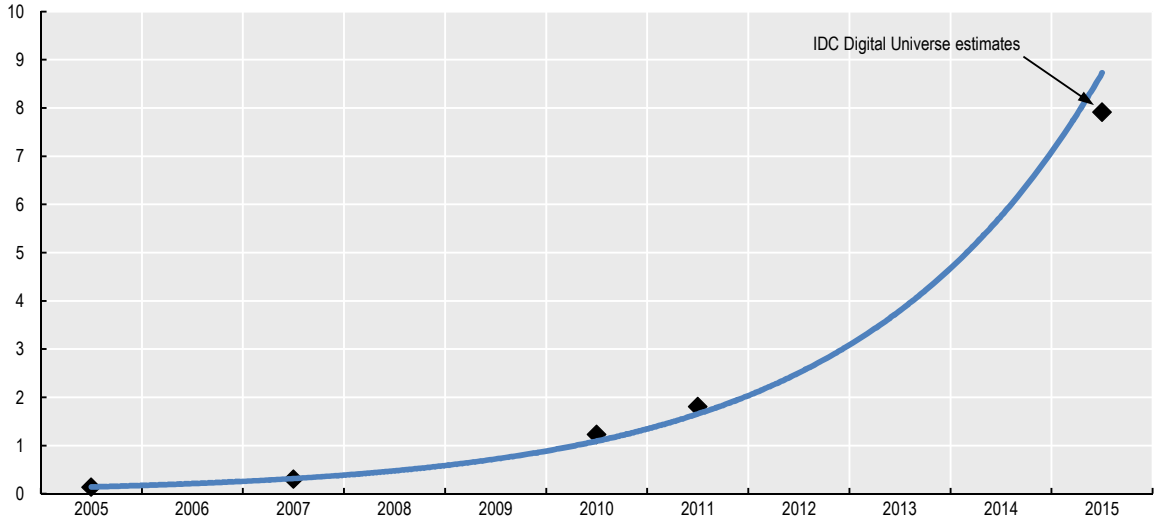
It's difficult to imagine the power that you're going to have when so many different sorts of data are available. (Berners-Lee, 2007)

Software is eating the world (Marc Andreessen, in Anderson, 2012)
… and the world is served in big chunks of data. (Esmeijer, Bakker, and de Munck, 2013)

More and more organisations are starting to leverage large volumes of (digital) data generated from myriad transactions and production and communication processes. These large streams of data, which are now commonly referred to as "big data", are generated through information and communication technologies (ICTs) including the Internet, as well as ubiquitous, wired sensors that are capturing activities in the physical world (see Chapter 3 of this volume). Measurement of the real total data generated, collected and stored is still speculative, but some sources suggest, for instance, that today more than 2.5 exabytes[1] (EB, a billion gigabytes) of data are generated every single day,[2] which is the equivalent of 167 000 times the information contained in all the books in the Library of Congress of the United States. This has led to an estimated cumulative data storage of around 8 zettabytes (ZB, a trillion gigabytes) in 2015 (Figure 1.1) and some estimates suggest that this will multiply by a factor of 40 by the end of this decade.[3] Today, the world's largest retail company, Walmart, already handles more than 1 million customer transactions every hour, which are imported into databases estimated to have contained more than 2.5 petabytes (PB, a million gigabytes) of data in 2010 (*The Economist*, 2010a).

Figure 1.1. **Estimated worldwide data storage**

In zettabytes (ZB, trillions of gigabytes)



Source: Based on the IDC (2012) Digital Universe research project.

The analysis of "big data", increasingly in real time, is driving knowledge and value creation across society; fostering new products, processes and markets; spurring entirely new business models; transforming most if not all sectors in OECD countries and partner economies; and thereby enhancing economic competitiveness and productivity growth. Algorithmic trading systems (ATS), for example, analyse massive amounts of market data on a millisecond basis to autonomously identify what to stock and when, and at what price to trade; this process was unheard of a decade ago (see Chapter 3). Traditional sectors such as manufacturing and agriculture are also being disrupted through the use of data and analytics, and are becoming more and more service-like (see Chapter 2). The German manufacturer of athletic shoes and sports equipment, Adidas, for instance, has redesigned many of its products as *data-driven services*, which are integrated via its online *miCoach* platform. This platform enables services related to physical activities

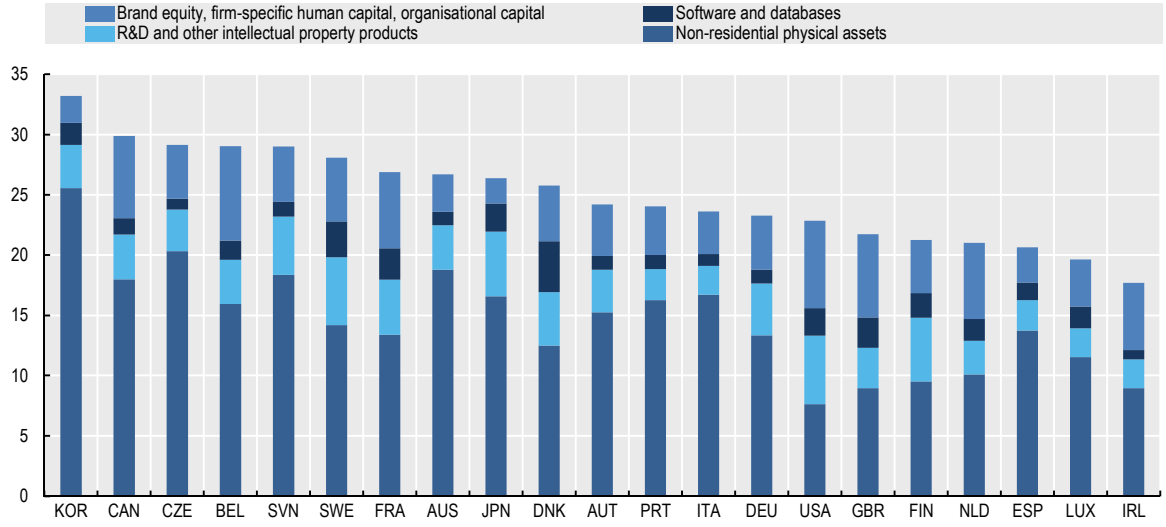
such as performance monitoring and training recommendation. In the public sector, the release of data as "open government data" can increase the transparency and accountability of governments, thus boosting public sector efficiency and public trust in governments (see Chapter 10). Better access to public sector information (PSI, including public sector data) can also empower entrepreneurs to develop new innovative commercial and social goods and services – such as the app "Asthmapolis", which is based on data released by the United States Government, and used to identify highly dangerous spots for asthmatic people. Since the app was created, hospitals in the United States have recorded a 25% decrease in asthmatic incidents.

The use of data and analytics to improve or foster new products, processes, organisational methods and markets – which is referred to hereafter as "data-driven innovation" (DDI) – is a new source of growth. It also represents a key opportunity for governments aiming to rebuild public trust through greater openness, transparency and accountability of the public sector. But governments need to address some major economic and societal challenges and risks in order to unleash the full potential of DDI and assure that its fruits contribute to the well-being of all citizens. These include, most prominently, the risk of i) barriers to the free flow of data (see Chapters 2, 3 and 4), ii) market concentration and competition barriers (Chapter 2), iii) base erosion and profit shifting (BEPS, Chapter 2), iv) privacy violation and discrimination (Chapter 5), v) dislocation effects in labour markets (Chapter 6), and with that vi) an emerging new digital or "data divide" that may hit developing economies particularly hard. Some of these challenges and risks deserve special attention from governments particularly concerned about social cohesion and rising inequality, which could hamper the economic resilience of their countries as stated by Ministers and Representatives[4] in the OECD (2014a) Ministerial Council Statement.[5]

DDI should be seen in a broader social and economic context in which knowledge-based capital (KBC) increasingly forms the foundation of 21st century knowledge economies, with data and software as one key pillar. In 2010, the OECD launched a horizontal project on *New Sources of Growth: Knowledge-Based Capital,* which provides evidence of the impact on growth, and the associated policy implications, of three main types of knowledge-based capital (KBC): i) computerised information (e.g. software and databases); ii) innovative property (e.g. patents, copyrights, designs and trademarks); and iii) economic competencies (e.g. brand equity, firm-specific human capital, networks of people and institutions, and organisational know-how) (OECD, 2013a).[6] The work highlighted that in some countries – such as Sweden, the United Kingdom and the United States – investment in KBC matches or exceeds investment in physical capital such as machinery, equipment and buildings (Figure 1.2). In many countries, such as Denmark, Ireland and Italy, business investment in KBC also rose higher as a share of GDP, or declined less, than investment in physical capital during the 2008-09 financial and economic crisis (OECD, 2013a).

Figure 1.2. **Investment in physical and knowledge-based capital, 2010**

As a percentage of value added of the business sector



Sources: *OECD Science, Technology and Industry Scoreboard 2013*, based on INTAN-Invest Database, www.intan-invest.net, and national estimates by researchers. Estimates of physical investment are based on OECD Annual System of National Accounts (SNA) and the INTAN-Invest Database, May 2013, http://dx.doi.org/10.1787/888932889820.

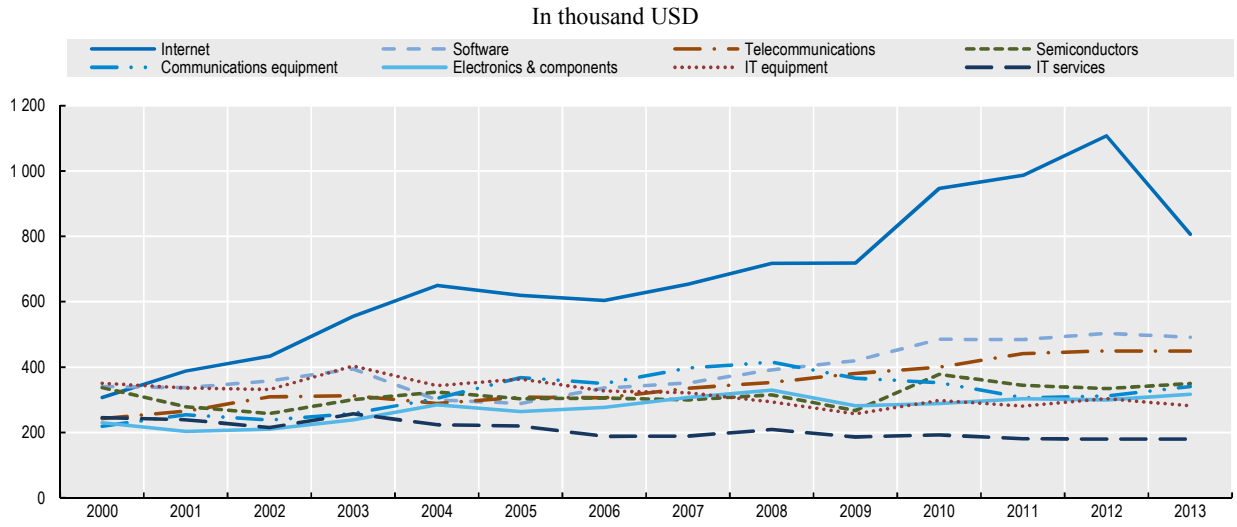
This synthesis chapter is structured as follows. It first presents available evidence on the increasing roles of data, analytics, and data-driven innovation. It then illustrates the context and policy issues related to the various aspects of DDI that are covered in this book, chapter by chapter. The discussion concludes by raising key challenges that most countries will face as DDI takes off and accelerates, and the policy considerations they will need to address.

1.1. The rise of "big data" and data-driven innovation

Leading the way: The ICT sector

ICT firms heavily rely on KBC investments, in particular software and data. This is especially apparent in the asset structure of Internet firms, such as Google and Facebook, where physical assets accounted for only around 15% of the firms' worth as of 31 December 2013.[7] Internet firms also enjoy huge productivity gains thanks to their KBC investments in software and data particularly. However, compared with other ICT firms, which also rely heavily on investments in software and data, Internet firms are by far more productive. Among the OECD area's top 250 ICT firms, Internet firms generated on average more than USD 1 million in revenues per employee in 2012 and more than USD 800 000 in 2013, while the other top ICT firms generated around USD 200 000 (IT services firms) to USD 500 000 (software firms) (Figure 1.3).

Figure 1.3. **Average revenue per employee of top 250 ICT firms, 2000-13**



Note: The presentation is based on averages for those firms reporting in 2000-13.

Sources: Based on OECD Information Technology database; compiled from annual reports, SEC filings and market financials, July 2014.

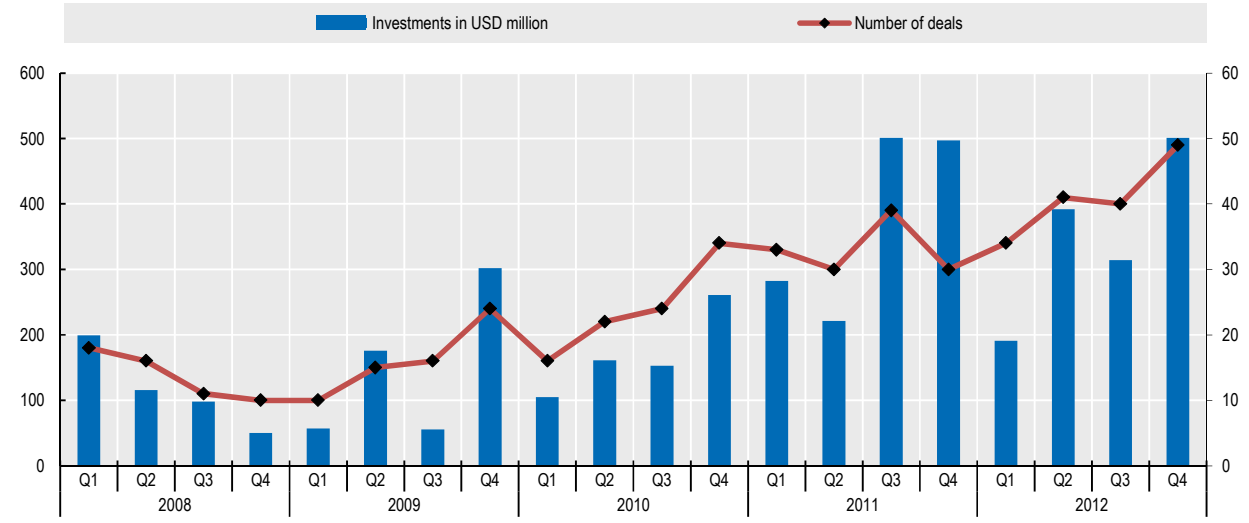
The business models of many Internet firms involve the collection and analysis of large streams of data collected from the Internet (OECD, 2012). By collecting and analysing "big data", a large share of which is provided by Internet users (consumers), Internet companies are able to automate their processes and to experiment with, and foster, new products and business models at much a faster rate than the rest of the industry. In particular, the advanced use of data and analytics enables Internet firms to scale their businesses at much lower costs than other ICT firms, a phenomenon that goes much further than what Brynjolfsson et al. (2008) describe as *scaling without mass*.[8]

The rest of the ICT sector (excluding Internet firms) has begun to recognise big data as a new business opportunity and is making significant investments to catch up and jump on the big data bandwagon. Estimates by IDC (2012) suggest that "big data technology and services" will grow from USD 3 billion in 2010 to USD 17 billion in 2015, which represents a compound annual growth rate (CAGR) of almost 40%. Technologies and services related to storage are expected to be the fastest growing segment, followed by networking and services, which explains the increasing role of IT equipment firms in this relatively new market.[9]

Top ICT companies are also strengthening their position through mergers and acquisitions (M&A) and/or through "co-opetition" (i.e. collaboration with potential and actual competitors). This includes in particular the acquisition of young start-ups specialised in big data technologies and services, and co-opetition via open source projects such as Hadoop (see Chapter 2). Data provided by Orrick (2012) on M&A deals (mainly in the United States) show that M&A activities have increased significantly since 2008 in terms of volume and number of deals: from 55 deals in 2008 to 164 in 2012, with almost USD 5 billion being invested over that period (Figure 1.4). In the first half of 2013 alone, big data companies raised almost USD 1.25 billion across 127 deals. IBM was the most active acquirer of big data companies in 2012, followed by Oracle.

Figure 1.4. **Big data-related financing activities, Q1 2008-Q4 2012**

Volume of investments in USD million (left scale) and number of deals (right scale)

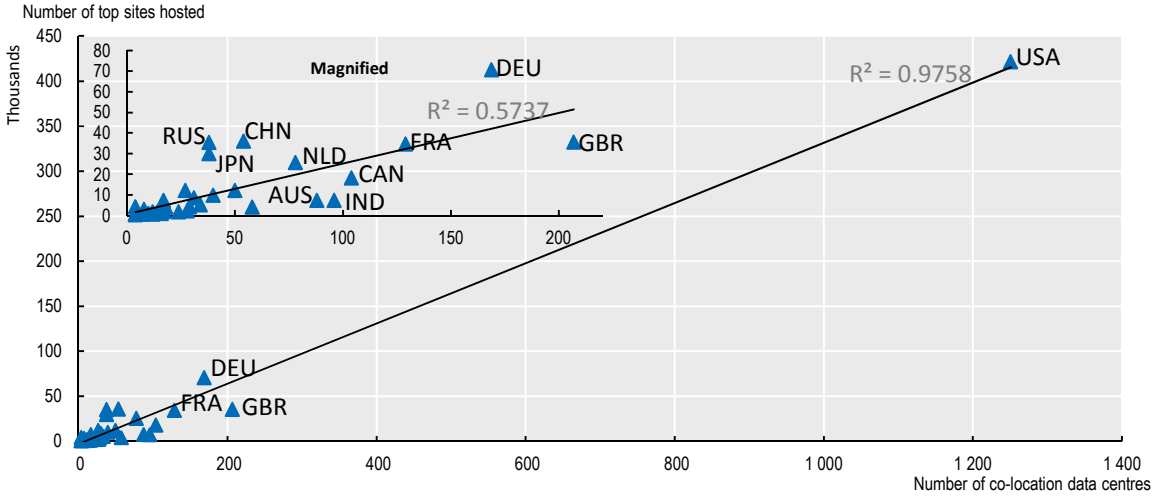


Source: Based on Orrick, 2012.

The combined effect of M&A, co-opetition, and the demand and supply of goods and services related to big data is the emergence of a *global data ecosystem* in which data and analytic services are traded and used across sectors and national borders (see Chapter 2 of this volume). The United States plays a central role, and countries such as Canada, Germany, France, Ireland, the Netherlands, Japan and the United Kingdom, as well as the People's Republic of China (hereafter "China"), India and Russia are catching up. The global data ecosystem involves global value chains (GVCs), in which companies increasingly divide up their data-related processes and locate productive activities in many countries. Figures on the distribution of data-driven services are not known. However, analysis of the world's top Internet sites suggests that data-driven services may be concentrated in the United States, which alone accounted for more than 50% of all top sites hosted in the OECD area, plus Brazil, China, Colombia, Egypt, India, Indonesia, Russia and South Africa in 2013 (Chapter 3). The number of top sites hosted correlates significantly with the number of co-location data centres (Figure 1.5).

Furthermore, top locations for data-driven services tend to be major exporters of ICT services (see Chapter 2). In 2013, the top ten exporters of ICT services were India, Ireland, the United States, Germany, the United Kingdom, China, France, the Netherlands, Belgium, and Sweden, and all – except Ireland, Belgium, and Sweden – are top locations for data-driven services. These countries are more likely to be the largest destination of cross-border data flows. As a consequence, the leading OECD area importers of ICT-related services are also the major sources for trade-related data; they include in particular the United States and Germany. (See Chapter 2 for further discussion on trade in data and ICT-related services.)

Figure 1.5. **Top locations by number of co-location data centres and top sites hosted**



Note: Number of top sites hosted based on analysis of 429 000 country code top-level domains (ccTLD) of the top one million sites collected in 2013. The remaining sites including the generic top-level domains were omitted from the list, as there are no reliable public data as to where the domains are registered.

Sources: Based on Pingdom, 2013; and www.datacentermap.com, accessed 27 May 2014.

Data-driven innovation across society

The economic impact of DDI goes far beyond market prospects for the ICT industry, although evidence strongly suggests that ICT firms are not only supplying products for data collection, processing and analysis, but also still leading in use of advanced data analytics. According to Tambe (2014), for example, only 30% of Hadoop investments come from non-ICT sectors, including in particular finance, transportation, utilities, retail, health care, pharmaceuticals and biotechnology firms. There is, however, a rapidly growing interest from non-ICT businesses across the economy in big data-related technologies and services to exploit data for innovation – that is to say, for developing new, or for improving existing, products, processes and markets (see Box 1.1 for the OECD definition of innovation).

Many organisations across the economy already benefit from significant investment in data in the form of traditional databases[10] for innovation, in particular in Finland, Denmark, Luxembourg, Sweden, the United Kingdom and the United States. As shown in Figure 1.2, investments in software and data (across the economy) accounted for an average share of slightly below 2% of business sector value added in OECD countries, with businesses in countries such as Denmark (4%), Sweden (3%), the United Kingdom and the United States (both 2%) leading in terms of the share of investment. With the exception of Sweden, these latter countries also saw a significant increase in software and data-related investments during the crisis, as did countries such as Luxembourg and Finland. Overall, investments in software have increased to 57% of total ICT investment in 2012, from less than 40% in 2000 (OECD, 2015a).

Box 1.1. **Defining innovation**

The latest (3rd) edition of the *Oslo Manual* defines innovation as the implementation of a new or significantly improved product (good or service), or process, new marketing method, or new organisational method in business practices, workplace organisation or external relations (OECD and Eurostat, 2005). This definition, for measurement purposes, captures the following four types of innovation:

- *Product innovation* – The introduction of a good or service that is new or significantly improved with respect to its characteristics or intended uses. This includes significant improvements in technical specifications, components and materials, incorporated software, user-friendliness and other functional characteristics.
- *Process innovation* – The implementation of a new or significantly improved production or delivery method. This includes significant changes in techniques, equipment and/or software.
- *Marketing innovation* – The implementation of a new marketing method involving significant changes in product design or packaging, product placement, product promotion or pricing.
- *Organisational innovation* – The implementation of a new organisational method in the firm's business practices, workplace organisation or external relations.

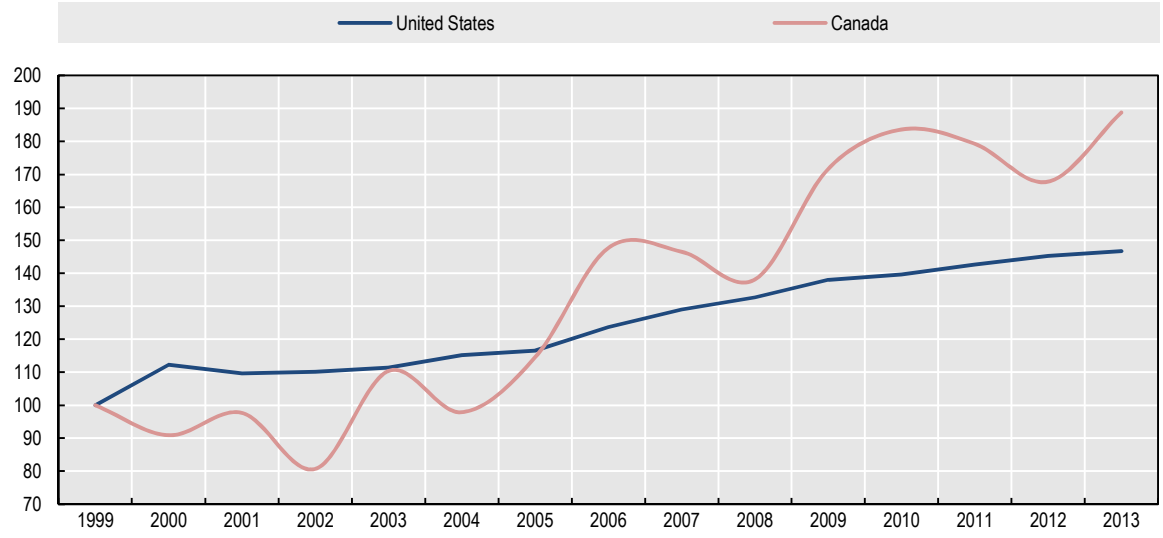
Source: OECD and Eurostat, 2005.

Increasing investments in software and databases go hand in hand with a growing data intensity of the economy as measured, for instance, by the share of data specialists in total employment. Employment figures for Canada and the United States show that the share of data specialists in total employment has increased since 1999 (Figure 1.6). The most data intensive industries employing the highest share of data specialists are still the ICT services industries, and in particular i) the IT and other information services industries, but also ii) insurance and finance, iii) science and research and development, iv) advertising and market research, as well as v) the public sector (see Chapter 6 of this volume). This is in line with findings by Tambe (2014) presented earlier and estimates by MGI (2011), according to which data intensity (measured as the average volume of data stored per organisation) is highest in financial services (including securities and investment services and banking), communication and media, utilities, government, and manufacturing. In these sectors, each organisation stored on average more than one petabyte (one million gigabytes) of data in 2009.

The following three sections describe the potential of DDI to contribute to productivity growth, well-being, inclusiveness and development. The process through which DDI creates value to achieve these policy objectives – the *data value cycle* – is presented in detail afterward.

Figure 1.6. **Trends in data intensity of the Canadian and United States economies, 1999-2013**

Index, share of data specialists[1] in total employment, 1999=100



Note: Data specialists do not correspond here to the 2008 International Standard Classification of Occupations (ISCO-08) definition presented in Box 6.4. in Chapter 6 of this volume. To be consistent across years, the definition has been slightly modified and does not include "Information security analysts" (SOC 2010 code 15-1122), "computer network architects" (15-1143) or "Computer occupations, nec" (15-1199) for the United States, and only include ISCO 08 code 212, "mathematicians, actuaries and statisticians", and (2521), "database designers and administrators" for Canada.

Sources: Occupational Employment Statistics (OES), US Bureau of Labor Statistics, www.bls.gov/oes/home.htm, November 2014; Statistics Canada, labour force survey, February 2015.

DDI can boost productivity growth

DDI is a disruptive new source of growth that could transform all sectors in the economy. Even traditional sectors such as retail, manufacturing and agriculture are being disrupted through DDI, as companies become more and more service-like, a trend that some have described using the term "servicification" (Lodefalk, 2010). Firms like Tesco, the UK supermarket chain, exploit huge data flows generated through their fidelity card programmes. The Tesco programme now counts more than 100 market baskets a second and 6 million transactions a day, and it very effectively transformed Tesco from a local, downmarket "pile 'em high, sell 'em cheap" retailer to a multinational, customer- and service-oriented one with broad appeal across social groups.

The world's largest company, Walmart, is even more progressive in its use of data and analytics. The company develops its own data analytic services via its subsidiary Walmart Labs, which is also actively contributing to the (co-)development of open source analytics. Walmart Labs' (internal) solution *Social Genome*, for example, allows Walmart to reach out to potential customers, including friends of direct customers, who have mentioned specific products online, to provide discounts on these exact products.[11] "This has resulted in a vast, constantly changing, up-to-date knowledge base with hundreds of millions of entities and relationships" (Big Data Startups, 2013).

In manufacturing, companies are increasingly using sensors mounted on production machines and delivered products to collect and process data on the machines' and products' operation, taking advantage of the Internet of Things (IoT) – the interconnection of "real world" objects. This trend, enabled by machine-to-machine

communication (M2M) and analysis of sensor data, has been described by some as "Industry 4.0" (Jasperneite, 2012), the "Industrial Internet" (Bruner, 2013), and "network manufacturing" (Economist Intelligence Unit, 2014). Sensor data are used to monitor and analyse the efficiency of products, to optimise their operations at a system-wide level, and for after-sale services, including preventive maintenance operations. The data are sometimes also used in collaboration with suppliers, and in some cases even commercialised as part of new services for existing and potential suppliers and customers.[12] For example, Germany-based Schmitz Cargobull, the world's largest truck body and trailer manufacturer, uses M2M and sensors to monitor the maintenance, travelling conditions and routes travelled by any of its trailers (Chick, Netessine and Huchzermeier, 2014; see also Vennewald, 2013). The insights generated by analysis of the data are used to help Schmitz Cargobull's customers minimise their usage breakdowns.[13] Quantitative evidence on the overall economic impact of DDI in manufacturing is still limited. Available estimates for Japan, for example, suggest that the use of big data and analytics by some divisions of Japanese manufacturing companies could lead to savings in maintenance costs worth almost JPY 5 trillion (which correspond to more than 15% of sales in 2010) and more than JPY 50 billion in electricity savings (MIC, 2013). For Germany it is estimated that Industry 4.0 can enable companies to boost their productivity by up to 30% (acatech, 2013), and to increase gross value added by a cumulative amount of up to EUR 267 billion by 2025 (BITKOM and Fraunhofer, 2014).

Agriculture is now being further modernised thanks to DDI, a development that is leading to huge productivity improvements and the reduction of environmental impacts. DDI in agriculture builds on geo-coded maps of agricultural fields and the real-time monitoring of every activity, from seeding to watering and fertilising, to harvesting. The data that are thereby generated can now be stored and analysed using cloud computing. As a result, farmers are today sitting on a wealth of agricultural data, which companies such as Monsanto, John Deere and DuPont Pioneer are trying to exploit through new data-driven goods and services (Noyes, 2014). John Deere, for example, is taking advantage of the "Industrial Internet" by integrating sensors into its latest equipment "to help farmers manage their fleet and to decrease downtime of their tractors as well as save on fuel" (Big Data Startups, 2013). The same sensor data are reused and linked with historical and real-time data on (e.g.) weather patterns, soil conditions, fertiliser usage and crop features, to optimise and predict agricultural production.[14] Traditional cultivation methods are thus improved and the wisdom and know-how of skilled farmers formalised. Overall, the use of data and analytics is estimated by some experts to improve yields by five to ten bushels per acre or around USD 100 per acre in increased profit (Noyes, 2014). This productivity increase comes at the right time, as the OECD and the Food and Agriculture Organization of the United Nations (OECD and FAO, 2012) call for a necessary food production increase by 60% for the world to be able to feed the growing population, which is expected to hit 9 billion in 2050.

There is as yet little evidence on the overall economic effects of DDI, but the few studies available suggest that firms using DDI raise labour productivity faster than non-users. A study of 330 companies in the United States by Brynjolfsson, Hitt and Kim (2011) estimates that the output and productivity of firms that adopt data-driven decision making are 5% to 6% higher than would be expected from their other investments in and use of ICTs. These firms also perform better in terms of asset utilisation, return on equity and market value. A similar study based on 500 firms in the United Kingdom by Bakhshi, Bravo-Biosca and Mateos-Garcia (2014) finds that businesses that make greater use of online customer and consumer data are 8% to 13% more productive as a result.[15] A recent

study by Tambe (2014) based on the analysis of 175 million LinkedIn user profiles, out of which employees with skills for big data-specific technologies have been identified, indicates that firms' investment in big data-specific technologies was associated with 3% faster productivity growth.[16] Overall, these studies suggest an approximately 5-10% faster productivity growth of DDI users compared to that of non-users.[17] However, it should be stressed that these estimates cannot be generalised, for a number of reasons. First, the estimated effects of DDI vary by sector and are subject to complementary factors, such as the availability of skills and competences and the availability and quality (i.e. relevance and timeliness) of the data used (see Chapter 4). But more importantly, these studies often suffer from selection bias, which makes it difficult to disentangle the effects of DDI from other factors at the firm level.[18] More comprehensive studies are therefore needed to better assess the impact of DDI on productivity growth.

DDI can contribute to well-being

The full impact of DDI goes beyond its positive effects on productivity growth. DDI can also contribute directly to the well-being of citizens, even if quantification of that contribution remains challenging because many if not most of the benefits related to the use of data are still not captured by market transactions (Mandel, 2012, 2013).[19] Citizens' use of open data as enabled by governments through their open data initiatives, for example, can increase the openness, transparency and accountability of government activities and thus boost public trust in governments. At the same time, it can enable an unlimited range of commercial and social services across society. For instance, "civic entrepreneurs" increasingly use available open data as promoted by the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* (PSI), in combination with other publicly available data sources, to develop apps that facilitate access to existing public services. Estimates on the economic impact of PSI (EUR 509 billion in 2008 for the reuse of PSI in the OECD area) focus on the commercial reuse of PSI and thus do not cover the full range of (social) benefits.

Science and education, health care services and public administration are the low hanging fruit policy makers can target in the relative short run to leverage DDI for growth and well-being. These sectors may be where adoption of DDI could have the highest impact. They employ the largest share of people who perform work related to the collection, processing and analysis of information and data. However, in these sectors, people are also still performing that work at a relatively low level of computerisation. In the United States, where data on working activities are available via the United States Department of Labor's O*NET system, almost 30% of the total employment in health care and social assistance, for instance, is in occupations largely involving information collection and analysis (e.g. records of patient medical histories, and test data or image analysis to inform diagnosis or treatment), but at the same time also involving a relatively low level of computer interaction.[20] Targeted promotion of the adoption of ICTs, and data and analytics in particular, could thus boost efficiency gains even further in these sectors.

In the area of science, the advent of new instruments and methods of data-intensive exploration could signal the arrival of new "data-intensive scientific discoveries", with new opportunities for knowledge creation. New instruments such as super colliders or telescopes, but also the Internet as a data collection tool, have been instrumental in these new developments in science, as they have changed the scale and granularity of the data being collected (see Chapter 7 of this volume). The Digital Sky Survey, for example, launched in 2000, collected more data through its telescope in its first week than had been amassed in the history of astronomy (The Economist, 2010a), and the new square

kilometre array (SKA) radio telescope could generate up to 1 petabyte (one million gigabyte) of data every 20 seconds (EC, 2010). Furthermore, the increasing power of data analytics has made it possible to extract insights from these very large data sets reasonably quickly. In genetics, for instance, DNA gene sequencing machines based on big data analytics can now read about 26 billion characters of the human genetic code in seconds. This goes hand in hand with the considerable fall in the cost of DNA sequencing over the past five years (see Chapter 3).

These recent developments in science obviously had significant impacts on health research and care, where the demographic evolution toward ageing societies and rising health costs are pressing for greater efficiency and for more responsive, patient-centric services (Chapter 8). At the core of DDI in the health sector are national health data, including but not limited to electronic health records and genetic, neuroimaging and epidemiological data. The efficient reuse of these data sets promises to improve the efficiency and quality of health care. In Finland for example, the content, quality and cost-effectiveness of treatment of a set of selected diseases are analysed by linking patient data across the whole cycle of care from admission to hospital, to care by their community doctor, to the medications prescribed and deaths (OECD, 2013c). The results of the analysis are made publicly available and have empowered patients and led to improvement in the quality of hospitals in Finland. In the particular case of the US health care system, MGI (2011) estimates that the use of data analytics throughout the system (clinical operations, payment and pricing of services, and R&D) could bring savings of more than USD 300 billion, two-thirds of which would come from reducing health care expenditures by 8%.[21]

New sources of data are already being considered by researchers who are seeking to improve research in and the treatment of diseases, as well as by individuals who are taking advantage of DDI to empower themselves for better prevention and care. For example, the social network PatientsLikeMe not only allows people with a medical condition to interact with and derive comfort and learn from other people with the same condition, but also provides an evidence base of personal data for analysis and a platform for linking patients with clinical trials. As another example, the so-called Quantified Self-movement has inspired its followers to use tools, like Fitbit, to track their every move and heartbeat, and to empower individuals to improve their health and overall well-being.

In the case of the public sector (intelligence and security excluded), there is some evidence of insufficient use of data that are generated and collected (see Chapter 10). According to MGI (2011), full use of data analytics in Europe's 23 largest governments may reduce administrative costs by 15% to 20%, creating the equivalent of EUR 150 billion to EUR 300 billion in new value and accelerating annual productivity growth by 0.5 percentage points over the next ten years.[22] The main benefits would be greater operational efficiency (due to greater transparency), increased tax collection (due to customised services, for example), fewer frauds and errors (due to automated data analytics). Similarly, a study of the United Kingdom shows that the public sector could save GBP 2 billion in fraud detection and generate GBP 4 billion through better performance management by using big data analytics (CEBR, 2012). Furthermore, data and analytics can be used to improve policy making by complementing official statistics (Reimsbach-Kounatze, 2015).

DDI can further inclusiveness and development

The potential of DDI to promote growth and contribute to well-being could provide a new opportunity to address the urgent needs of developing economies (Gerdon and Reimsbach-Kounatze, 2015). Increasingly, a wide range of data sources, including mobile phones, social media and the public sector, are being explored by governments, businesses, researchers and citizens groups and used to foster development (UN Global Pulse, 2012; WEF, 2012). International initiatives have formed that investigate the capabilities of data analytics for development. *Paris21*, the Partnership in Statistics for Development in the 21st century, brings together users and producers of statistics in developing and developed countries to strengthen statistical capacities and promote the use of reliable data (Letouzé and Jütting, 2014). Meanwhile, the United Nations (UN) Global Pulse initiative was launched by the Executive Office of the UN Secretary-General in response to the need for more timely data to track and monitor the impacts of global and local socio-economic crises (UN Global Pulse, 2012). The UN, moreover, announced the need for a data revolution for a future development agenda beyond 2015 to succeed the United Nations Millennium Development Goals. The Harvard Humanitarian Initiative, the MIT Media Lab and the Overseas Development Institute jointly formed the Data-Pop Alliance to work on big data for development to improve decisions and empower people in a way that avoids the pitfalls of a new digital divide, de-humanisation and de-democratisation (Letouzé and Jütting, 2014).

Significant progress has been made with the use of data analytics for crisis prevention and disaster management (see Box 1.2).[23] Thailand, for instance, is monitoring natural disaster-prone areas such as the coastline, rivers and forests with satellite and ground sensors in order to better react in emergency situations. The Kenyan-based non-profit software company Ushahidi created a system to collect real-time data from eyewitnesses of violence in the aftermath of Kenya's disputed 2007 presidential election; the system has since been used to gain a better understanding of complex situations such as the 2010 earthquake in Haiti, the Syrian Conflict beginning in 2011, and the Ebola epidemic in 2014. Recently, UN Global Pulse has focused on identifying and quantifying discussion themes in Twitter data in order to investigate how people cope in crisis situations such as food price crises or economic crises (UN Global Pulse, 2014).

Box 1.2. **Big data for disaster management**

Real-time analysis of a wide range of data generated through social media, mobile devices and physical sensors (e.g. the Internet of Things) provides a new opportunity for addressing complex societal challenges, including in particular crisis prevention and disaster management. A series of documentary films, "Disaster Big Data", produced by Japanese public broadcaster NHK has shown how data analytics can help build a better understanding and improve response to tremendous disasters such as the one caused by the 2011 earthquake and tsunami in Japan.[1] Data and analytics, together with other ICTs, play an important role at every stage, from prediction to incident management to reconstruction. M2M communication, for instance, can enable the collection of data from water-level sensors, mudslide sensors and GPS sensors for a real-time monitoring and alert system. Japan is now introducing just such an advanced disaster management system, combining this information with location data.

1. See www.nhk.or.jp/datajournalism/about/index_en.html, accessed 15 May 2015.

DDI for development could provide some countries with the capacity to "leapfrog" in critical development areas such as transport, finance and agriculture. In transport, the use of data and analytics could improve transport systems in mega cities (see Chapter 9). The online platform Tsaboin, for instance, crowdsources[24] traffic data based on passenger information around bus stops in Lagos, Nigeria, where no official traffic feed exists, to enable users to check the traffic information in real time and make "smarter" traffic decisions.[25] In finance, Cignifi is used to develop mobile-based credit scores. This start-up mines cellphone data to assign a credit score to unbanked potential clients.[26] In the field of agriculture, data analytics can improve the work of farmers through information, forecasting and evaluation, particularly on the local level. The International Center for Tropical Agriculture (CIAT) developed a climate-smart, site-specific recommendation engine for Colombian rice farmers, based on meteorological data and seasonal forecasts.

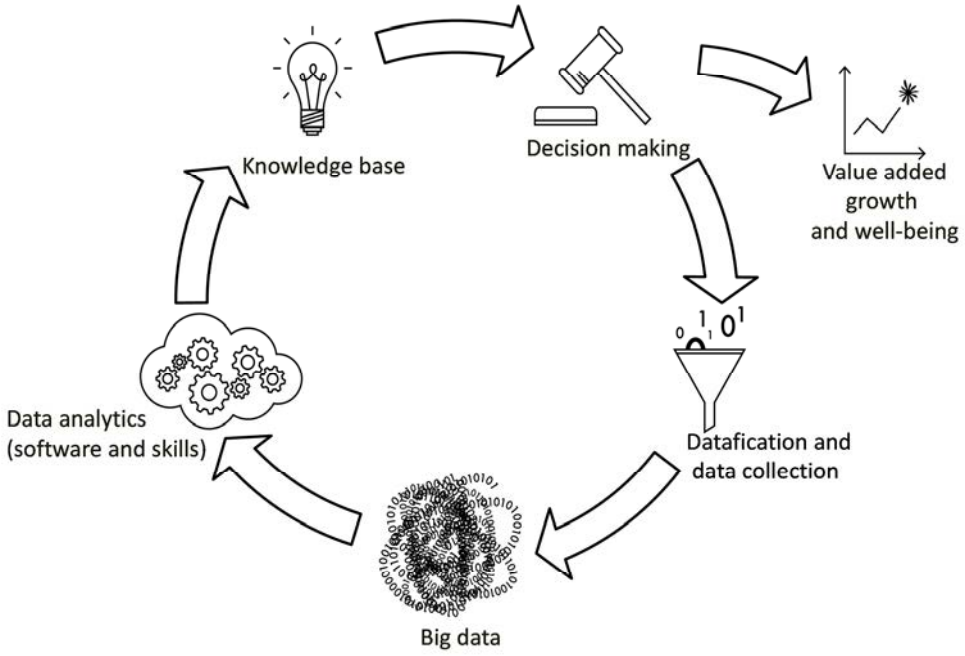
The data value cycle: From datafication to data analytics and decision making

Policy and decision makers aiming at leveraging DDI for growth, well-being and development must understand the process through which data are transformed to finally lead to innovation. In this volume, DDI is described as a sequence of phases from datafication to data analytics and decision making. This process, however, is not a (linear) value chain, but a *value cycle* that involves feedback loops at several phases of the value creation process. The stylised data value cycle illustrated in Figure 1.7 includes the following phases:

- *Datafication and data collection* – These refer to the activity of data generation through the digitisation of content, and monitoring of activities, including real-world (offline) activities and phenomena, through sensors.
- *Big data* – This refers to the result of datafication and data collection that together lead to a large pool of data that can be exploited through data analytics.
- *Data analytics* – Until processed and interpreted via data analytics, big data are typically useless since the first glance reveals no obvious information. Data analytics is increasingly undertaken via cloud computing.
- *The knowledge base* – This refers to the knowledge that is accumulated through learning over time. Where machine learning is involved, the knowledge base reflects the state of the learning system. The knowledge base is the "crown jewels" of data-driven organisations, and therefore enjoys particular protection through legal (e.g. trade secrets – see OECD, 2015b) and technical means (see Chapter 5 on the implications for digital risk management).
- *Data-driven decision making* – The value of data is mainly reaped at two moments: first when data are transformed into knowledge (gaining insights), and then when they are used for decision making (taking action). Decisions taken can in turn lead to more or different data generated and thus trigger a new data value cycle.

Analytics and the value cycle are discussed further below, in the section titled "How data now drive innovation – The focus of Chapter 3".

Figure 1.7. **The data value cycle**



1.2. Objectives and structure of this volume

This volume includes ten chapters discussing the various key aspects of DDI with the aims to: i) improve the evidence base on the role of DDI for promoting growth and well-being, and ii) provide policy guidance on how to maximise the benefits of DDI, and mitigate the associated economic and societal risks. The insights it presents are intended to assist policy makers in better understanding DDI and in incorporating its multidimensionality into policy design. This will, according to the OECD (2014a) Ministerial Council Statement, "help identify trade-offs, complementarities[27] and unintended consequences of policy choices", in line with the common goal of building and maintaining "resilient economies and inclusive societies". These insights can also feed a wide range of future OECD work, including the preparations for the 2016 meeting at Ministerial level organised by the OECD Committee on Digital Economy Policy (CDEP) as well as the current revision of major OECD instruments related to data access, linkage and reuse. These include the OECD (2008) *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* of 30 April 2008 and the OECD (2006) *Recommendation of the Council concerning Access to Research Data from Public Funding* of 14 December 2006.

The remainder of this section introduces the context and policy issues related to the various aspects of DDI that the reader will encounter in this volume.

Mapping the global data ecosystem – The focus of Chapter 2

For many of the steps in the value creation process along the data value cycle presented above (Figure 1.7), organisations will have to involve third parties around the world, because they lack the experience, technological resources and/or talent to deal with the multidisciplinary aspects of data and analytics on their own. The resulting global value chain (GVC) is in most cases specifically tailored towards the goal that is being

pursued. What emerges from all this interaction is a global data ecosystem in which, more than ever before, data and analytic services are traded and used across sectors and national borders. The concept of an ecological approach to describe business environments was introduced by Moore (1993) to describe how companies should not be viewed as members of a single industry "[…] but as part of a business ecosystem that crosses a variety of industries." In these ecosystems, collaborative arrangements of firms combine their individual offerings to create coherent, customer-tailored solutions (Adner, 2006).

Key actors in the ecosystem

The global data ecosystem is evolving swiftly due to the increasing number actors, many of which typically have multiple roles, goods and services, technologies, and business models.[28] The data ecosystem is seen in this report as a combination of layers corresponding to key roles of actors, where the underlying layers provide goods and services to the upper layers.

A first layer of actors includes Internet service providers; these form the backbone of the data ecosystem through which data is exchanged. A second layer includes IT (hardware and software) infrastructure providers that offer data management and analysis tools and critical computing resources – including, but not limited to, data storage servers, database management and analytic software, and (most importantly) cloud computing resources. The third layer includes data (service) providers: i) data brokers and data marketplaces that commercialise data across the economy; ii) the public sector with its open data initiatives (see Chapter 10); and iii) consumers, which actively contribute their data to the data ecosystem through new services provided by innovative businesses and through data portability initiatives. A fourth layer includes data analytic service providers – businesses that provide data aggregation and analytic services, mainly to business customers. Finally, there are data-driven entrepreneurs that build their innovative businesses based on data and analytics available in the data ecosystem. DDI from these entrepreneurs can be applied to science and research (see Chapter 7), health care (Chapter 8), and smart cities (Chapter 9), and public service delivery (Chapter 10).

Interactions in the ecosystem

Interaction among the actors that structure the data ecosystem could best be described as "co-opetition", a combination of competition and collaboration. As with many innovation ecosystems, collaboration among individual companies allows them to create value that no single company can deliver on its own. Promising (and often specialist) start-ups emerge, which are eventually acquired by larger companies wishing to improve and augment their propositions with analytics platforms, visualisations and applications (ESG, 2012). In the past five years the focus on mergers and acquisitions, in terms of both deals and (especially) investments, has shifted from big data infrastructure to big data analytics and applications.

Recent years have also seen the emergence of "data markets" – online services that host data from various publishers and offer the (possibly enhanced) data to interested parties (Dumbill, 2012). One important distinguishing factor between data brokers and data market providers is that data brokers are actively engaged in the collection of additional data and their aggregation, while data market providers are intermediaries through which data controllers (including brokers) can offer their data sets.

The data ecosystem's value chains are truly global; companies increasingly divide up their production processes and locate productive activities in many countries. Data may be collected from consumers or devices located in one country through devices and apps developed in another country. They may then be processed in a third country and used to improve marketing to the consumer in the first country and/or to other consumers around the globe. Many global value chain activities are captured in international trade, not only in ICT services provided by actors in the IT infrastructure layer, but also in other data-intensive services such as finance, e-commerce, and research. In fact the leading OECD importers of ICT-related services are also the major sources for trade-related data.

Key challenges in the global data ecosystem

The globally distributed nature of the data ecosystem, its extreme interconnectedness, and the interdependencies of its actors and their technologies and resources raise a number of policy issues. One such challenge is the difficulty of value attribution: this challenges measurement but also taxation policies. A number of governments have raised concerns that characteristics of the global data ecosystem could create opportunities for base erosion and profit shifting (BEPS) through "aggressive tax planning by multinational enterprises making use of gaps in the interaction of different tax systems to artificially reduce taxable income or shift profits to low-tax jurisdictions in which little or no economic activity is performed" (OECD, 2014b). A second concerns the key points of control and competition: some dominant actors in the data ecosystem may have significant control and power over certain activities through which the data ecosystem could be shaped, and eventually disrupted. The third involves the free flow of data, which favours global competition among actors of the data ecosystem. Barriers to the flow can limit the effects of DDI, by limiting for example trade and competition. Finally, there remain barriers to data interoperability – especially in sectors requiring significant investment, with a high threshold for new entrants – and portability, which refers to the capacity of reusing data for new applications.

The following four chapters focus on the key issues that decision and policy makers need to consider in more detail, including:

1. the key factors through which decision and policy makers can leverage DDI for growth and well-being: (i) "How data now drive innovation – The focus of Chapter 3", and (ii) "Drawing value from data as an infrastructure – The focus of Chapter 4"
2. the two major policy challenges decision and policy makers need to address to mitigate the economic and societal risks that come with DDI: (i) "Building trust for data-driven innovation – The focus of Chapter 5", and (ii) "Skills and employment in a data-driven economy – The focus of Chapter 6".

How data now drive innovation – The focus of Chapter 3

While the importance of data, both economically and socially, is not new, a confluence of three major socio-economic and technological trends along the data value cycle (Figure 1.7) is making DDI a new phenomenon today and a new source of growth.

The enablers of data-driven innovation

The first is the exponential growth in data generated and collected, driven by high-speed mobile broadband, and the Internet of Things, including sensors and sensor

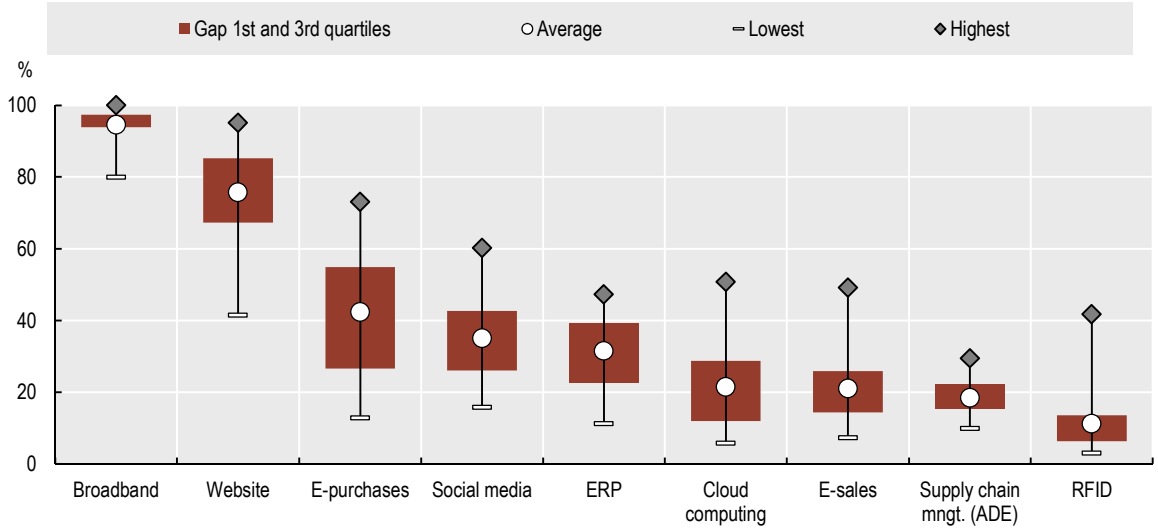
networks enabling the ubiquitous "datafication" of the physical world, and machine-to-machine communication (M2M) empowering data exchange in that world. It is estimated that the average number of Internet-connected devices per household in OECD countries – which today totals ten for an average family of four persons (including two teenagers) – could reach 50 by 2022. One question that arises is whether networks will be able to support all the devices that will be coming on line (an estimated 50 billion devices by 2025).

The second major development favouring DDI is the pervasive power of data analytics, which is now becoming affordable for start-ups and small and medium-sized enterprises (SMEs). The huge volume of data generated by the Internet has no value if no information can be extracted from the data; data analytics refers to a set of techniques and tools that are used to extract information from data, revealing the context in which the data are embedded and their organisation and structure. It reveals the "signal from the noise" – patterns, correlations and interactions among all the pieces of information. The adoption of data analytics has been greatly facilitated by the declining cost of data storage and processing. Cloud computing has been key to this cost reduction. There remain, however, significant issues limiting adoption of cloud computing, which is still used to a much less degree than the high level of broadband connectivity and website adoption would suggest (Figure 1.8). Besides a low capacity to change of many businesses (see Chapter 6), privacy and security (Chapter 5) are among the two most pressing issues limiting cloud computing adoption. Another major challenge is the lack of appropriate standards and the potential for vendor lock-in due to the use of proprietary solutions: applications developed for one platform often cannot be easily migrated to another application provider.

The third factor is the emergence of a paradigm shift in knowledge creation and decision making. Those two moments – when data are transformed into information and knowledge (gaining insights) and then used for decision making (taking action) – are when the social and economic value of data is mainly reaped. Separating these concepts (data, information, and knowledge) is important to better understand data-driven value creation. The distinction can help explain how it is one can have a lot of data but not be able to extract value from them when not equipped with the appropriate analytic capacities; and how one can have a lot of information (extracted from data), but not be able to gain knowledge from it, a phenomenon nowadays better known as "information overload". Data analytics today can help gain insights through i) extracting information from unstructured data (i.e. that lack a predefined data mode); ii) real-time monitoring; and iii) inference – the "discovery" of information even if there was no prior record of such information, through "mining" available data for patterns and correlations – and prediction.

Figure 1.8. **The diffusion of selected ICT tools and activities in enterprises, 2013**

Percentage of enterprises with ten or more persons employed



Note: For countries in the European Statistical System, sector coverage consists of all activities in manufacturing and non-financial market services, and data on e-purchases and e-sales refer to 2013. For Australia, data refer to the fiscal year 2013/14, ending on 30 June and include agriculture, forestry and fishing activities. For Canada and Japan, data refer to 2013 except cloud computing (2012). For Korea, data refer to 2013. For Mexico, data refer to 2012 and to establishments with 10 or more persons employed. For New Zealand, data refer to the fiscal year 2013/14, ending on 31 March. For Switzerland, data refer to 2011.

Source: OECD (2015c).

The implications for the quality of decision-making

Data now have an even bigger role in the decision-making process than in the past. Three major trends account for this: (i) Human decision making is increasingly based on rapid data-driven experiments; (ii) crowdsourcing – "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community" (Merriam-Webster, 2014) – has been made further affordable; and (iii) decision-making is increasingly being automated thanks to advances in artificial intelligence. In fact, one of the largest impacts of data on (labour) productivity is expected to come from decision automation, thanks to "smart" applications that are "able to learn from previous situations and to communicate the results of these situations to other devices and users" (OECD, 2013d).

As a result of these three major trends, analytics obviates the need for decision makers to understand the phenomenon before they act on it: in other words, first comes the analytical fact; then the action; and last, if at all, the understanding.[29] This can raise serious issues, in particular because the use of data analytics does not come without limitations. There are considerable risks that the underlying data and analytic algorithms could lead to unexpected (false) results, a risk heightened when decision making is automated. Three types of errors could occur: (i) those due to poor-quality data (which will almost always lead to poor results); (ii) those that come with inappropriate use of data and analytics (there will be wrong results if the data used are irrelevant and do not fit the business or scientific questions they are supposed to answer); and (iii) those caused by unexpected changes in the data environment. These last may be intentional; sometimes analytics can be easily "gamed" once the factors affecting the underlying algorithms have

been understood. Or the wrong results may not be intentional and due to constant changes in the data environment; patterns in the data collected are therefore hardly robust over time.

Drawing value from data as an infrastructure – The focus of Chapter 4

Data has become a key infrastructure for 21st century knowledge economies. Data are not the "new oil" as still too often proclaimed. They are rather an infrastructure and capital good that can be used across society for a theoretically unlimited range of productive purposes, without being depleted.

Open data, data commons, and data philanthropy

Data provide economies with significant growth opportunities through spillover effects in the support of the downstream production of goods (including public and social goods).[30] And as with any infrastructure, there can be significant (social) opportunity costs in limiting access. Open (closed) access enables (restricts) user opportunities and degrees of freedom in the downstream production of private, public and social goods (Frischmann, 2012). Especially in environments characterised by high uncertainty, complexity and dynamic changes, open access can be an optimal (private and social) strategy for maximising the benefits of an infrastructure. Data markets may not be able to fully serve social demand for data if there is a demand manifestation problem – as there certainly can be – in the data ecosystem. In addition, the context dependency of data and the dynamic environment in which some data are used (e.g. research) make it almost impossible to fully evaluate ex ante the potential of data, and would exacerbate a demand manifestation problem.

This calls for governing data through non-discriminatory access regimes and commons (see Frischmann, Madison and Strandburg, 2014). In contrast to Hardin's (1968) "tragedy of the commons", where free riding on common (natural) resources leads to degradation and depletion of resources, the "comedy of the commons" (Rose, 1986) – where greater social value is created with greater use of common resources – is possible in the case of non-rivalrous goods such as data. This is the strongest rationale for policy makers to promote access to data, either through "open data" in the public sector, "data commons" such as in science, or through the more restrictive concept of "data portability" to empower consumers. The accumulation of data does come with certain costs (e.g. storage) and risks (e.g. privacy violation and digital security risks). Nevertheless, the advantages for individuals and businesses are clear.

Most definitions for open data point to a number of criteria or "principles". According to the OECD (2006) *Council Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*, for example, openness means i) access that should be granted on equal or non-discriminatory terms, and ii) access costs that should not exceed the marginal cost of dissemination. And it is important to note that the concept of open data is not limited to science (Chapter 7) or the public sector (Chapter 10). For instance, "data philanthropy", whereby the private sector shares data both to enable societal benefits such as by supporting more timely and targeted policy action for development.

Towards a common data governance framework

Among the criteria listed in the many definitions of open data, non-discriminatory access (or "access on equal terms", as stated in the OECD [2006] Recommendation) is

central. Access independent of identity and intent can be crucial for maximising the value of data across society, as it keeps the range of opportunities as wide as possible. Three factors in particular affect the level of non-discriminatory access.

One is the data's technological design – they need to be made available, ideally on line; machine readable, i.e. structured; and linkable. Intellectual property rights (IPRs) are a second factor, for they can limit or prevent the (re-)use and distribution of open data. Some open data initiatives therefore explicitly state that open data should be free of any IPRs, although in other cases innovative IP regimes are used and even promoted through open data regimes, as long as they do not restrict the rights of users to reuse and sometimes redistribute the data.

Pricing, the third factor, will have less of an impact on the degree of openness than technological design or IPRs, but it can still be one of the most challenging factors, because optimal pricing can be hard to determine. Many governments wish to engage in cost recovery, partly for budgetary reasons and partly based on the principle that those who benefit should pay. But the calculation of the overall benefits can be problematic due to significant spillover effects through the creation of public and social goods based on open data. Furthermore, as Stiglitz et al. (2000) have argued, if government provision of a data-related service is a valid role, generating revenue from that service is not. Many open data initiatives therefore encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" as stated in the OECD (2005) Recommendation.

Pricing is challenging mainly due to the fact that data have no intrinsic value, as the value depends on the context of their use. A number of factors can affect that value, in particular the accuracy and the timeliness of data. The more relevant and accurate data are for the particular context in which they are used, the more useful and thus valuable data will be. This of course implies that the value of data can perish over time, depreciating as they become less relevant for their intended use. There is thus a temporal premium that is motivated by the "real-time" supply of data, for example in the financial sector.

Better data governance regimes are needed to overcome barriers to data access, sharing and interoperability. These regimes can have an impact on the incentives to share and the possibility of data being used in interoperable ways. The elements to consider for effective data governance include data access and reuse; portability and interoperability; linkage and integration; quality and curation; "ownership" and control; and value and pricing.

Ownership is singled out, because it a questionable appellation when it comes to data and personal data in particular. In contrast to other intangibles, data typically involve complex assignments of different rights across different data stakeholders. Those different stakeholders will typically have different power over the data, depending on their role. In cases where the data are considered "personal data", the concept of data ownership by the party that collects personal data is even less practical since privacy regimes grant certain explicit control rights to the data subject, as for example specified by the Individual Participation Principle of the OECD (2013e) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*.

Building trust for data-driven innovation – The focus of Chapter 5

Critical to reaping the substantial economic benefits of DDI – as well as to realising the full social and cultural potential of that innovation – is trust. Trust is a complex issue,

and yet there is consensus that it plays a central if not vital role in social and economic interactions and institutions. Trust is seen as central for efficiency gains realised thanks to the reduction of transaction costs in social and economic interactions. In reducing transaction costs and frictions, trust generates efficiency gains. Trust is therefore considered by some to be a "social capital" and a determinant of economic growth, development, and well-being. The OECD (2011) provides quantitative evidence that high country trust is strongly associated with high household income levels. While trust can be built, it can also erode over time if overexploited as discussion on the recent financial crisis (Allen, 2013; OECD, 2013f) and the revelations about intelligence gathering (Croft, 2014; Naughton, 2015) have suggested.[31] The main components of *trust in the digital economy* are security, privacy and consumer protection.

From traditional security to digital security risk management

DDI relies on an intricate, hyper-connected ICT environment in which security threats have changed in both scale and kind. They include organised crime groups, "hacktivists", foreign governments, terrorists, individual "hackers" – and sometimes, business competitors. There are in addition the non-intentional digital threats, such as hardware failure and natural disasters.

Many stakeholders continue to adopt a *traditional security approach* that not only falls short of appropriately protecting assets in the current digital environment, but also is likely to stifle innovation and growth. That traditional approach aims to create a digital environment secure from threats that can undermine the "AIC triad": data's availability (accessibility and usability upon demand by an authorised entity); integrity (quality in terms of accuracy and completeness); and/or confidentiality (prevention of data disclosure to unauthorised individuals, entities or processes). To preserve each of these dimensions, security experts put in place "controls", "mechanisms" or "safeguards", generally based on technologies, that form a perimeter around the protected assets to secure them.

The problem here is that data-intensive economic and social activities introduce a level of complexity to the point where the traditional security approach cannot scale up. First, these data-intensive activities rely on information systems and networks to become more open and interconnected, enabling data flows to be exchanged easily, flexibly and cheaply, with a potentially unlimited number of partners outside the perimeter. Second, DDI relies on the capacity to exploit the dynamic nature of the digital environment – rapidly connecting, matching and analysing what was previously not related in order to create new assets. Third, traditional security can deal with increased volumes and diversity if the data are located within that defined perimeter and their processing is not subject to continuously unpredictable uses and flows. However, the uncertainty already introduced by the open and dynamic nature of DDI grows, sometimes exponentially, with these increases.

As a result, the traditional security approach, which can only operate at the cost of reducing complexity and increasing stability, will inevitably slow innovative usage and, ultimately, undermine the economic and social benefits of interoperable ICTs.

With the risk-based management approach, the value of data-intensive activities is not limited to the digital storage and processing of a large quantity of data ("big data"), but rather to the capacity to manage a data value cycle (Figure 1.7). The objective of digital security risk management is therefore to increase the likelihood of economic and social benefits from the data value cycle by minimising potential adverse effects of uncertainty

related to the availability, integrity and confidentiality of the cycle (the AIC triad). Unlike the traditional security approach, digital security risk management does not aim to create a secure digital environment to eliminate risk. Instead, it creates a framework to select proportionate and efficient AIC security measures in light of the benefits expected from the cycle.

That raises the key question of responsibility. Traditional security focuses on securing the digital environment. Therefore, in most cases, the party responsible for the provision of the environment (generally the IT department) takes responsibility for its security, and users of the environment do not have to be concerned with it. In contrast, from a digital security risk management perspective, responsibility cannot be delegated to a separate party. Managing risk means accepting a certain level of risk – or deciding not to accept it, and therefore not to realise the benefits. The primary responsibility for managing risk should therefore mirror the responsibility for achieving the objectives and realising the benefits (leadership).

Privacy protection for data-driven innovation

Each step of the data value cycle (Figure 1.7) on which data-driven innovation relies can raise privacy concerns. *Step 1* is the initial data collection, which is becoming increasingly comprehensive, diminishing an individual's private space. Some of the data collected is *volunteered* and thus knowingly and willingly provided by the individual as it is often essential to the completion of an online transaction. An increasing share of data in contrast is *observed*, based on the online tracking of individuals and the collection and analysis of related personal information.

Step 2 is the massive storage of data, which increases the potential of data theft or misuse by malicious actors and other consequences of a data security breach, the risks of which may not be easy to ascertain. Where personal data are collected, stored or processed, security incidents can heavily affect individuals' privacy as high-profile *data breaches*[32] have demonstrated. Cyber-attacks still remain the most frequent cause for data breaches in terms of records stolen but not in number of incidents. These incidents come along with significant costs to individuals but also to the firms suffering the data breaches.

Steps 3 and 4 involve inferences of information and knowledge enabled by data analytics, which often go well beyond the data knowingly provided by a data subject, diminishing an individual's control and creating information asymmetry. Advances in data analytics, make it increasingly easy to generate *interferences* from data collected in different contexts, even if individuals never directly shared this information with anyone. Once *linked* with sufficient other information, data analysts can predict, with varying degrees of certainty, the likelihood that an individual will possess certain characteristics, building a profile. This increased capacity of data analytics is illustrated by Duhigg (2012) and Hill (2012), who describe how the United States based retailing company Target "figured out a teen girl was pregnant before her father did" based on specific signals in historical buying data.[33]

Finally, data-driven decision making (*Step 5*) can lead to a real-world (discriminatory) impact on individuals and other harms. Concerns have been raised that the information inferred through data analytics could be used to exploit the vulnerabilities and receptiveness of individuals in a way that not only induces them to undertake certain actions (e.g. purchase products), but that alters their preferences for these actions. In addition, private actors increasingly rely on the predictive capabilities of data analytics in

their search for competitive advantage. While these predictive analyses may result in greater efficiencies, they may also perpetuate existing stereotypes, limiting an individual's ability to escape the impact of pre-existing socio-economic indicators. A well-known example in this regard is "price discrimination" where firms are selling the same good to different customers for different prices, even though the cost of producing for the two customers is the same. Certain uses of data analytics may also have more serious implications for individuals, for example, by affecting their ability to secure employment, insurance or credit, and this is the more severe when decision-making processes are fully automated.

Data analytics may thus impact core societal values such as individuals' liberty, when for instance creating a "chilling effect" in which an individual curtails communications and activities in fear of uncertain but possibly adverse consequences, or a "filter bubble" (Pariser, 2012) which narrows the range of views exposed to an individual as a result of efforts to personalise content and other products and services. Where the issue of information asymmetry is further exacerbated by the limited transparency of data analytics, individuals will remain unaware that data analytics is affecting their decision making and even preferences, and they will have considerable difficulty ascertaining how exactly analytics is being used to influence them.

There have been several policy responses to improve the effectiveness of privacy protections in the context of DDI. One set of initiatives is grouped under a heading of improving transparency, access and empowerment for individuals. An element in a number of these initiatives is data portability, which allows users to more easily change data controllers by reducing switching costs, and enables them to analyse their own data for their own benefit by receiving it in a usable format (see also Chapter 4). Another emerging element includes the means through which the transparency of the processes and algorithms underlying data analytics (i.e. *algorithmic transparency*)[34] can be increased (see Box 1.3, sell also Annex on the highlights of the *2014 OECD Global Forum on the Knowledge Economy*). A second area of focus is the promotion of responsible usage of personal data by organisations. The promise of technologies used in the service of privacy protection has been long noted as another area. Finally, application of risk management to privacy protection is highlighted as providing another possible avenue. Perhaps the most difficult policy prescription is a need for greater effort to articulate substantive boundaries within which responsible uses of data and analytics would be limited, including the boundaries within which fully automated decision-making would be appropriate. Determining where these boundaries lie – and who should make this determination – will become an increasingly necessary task.

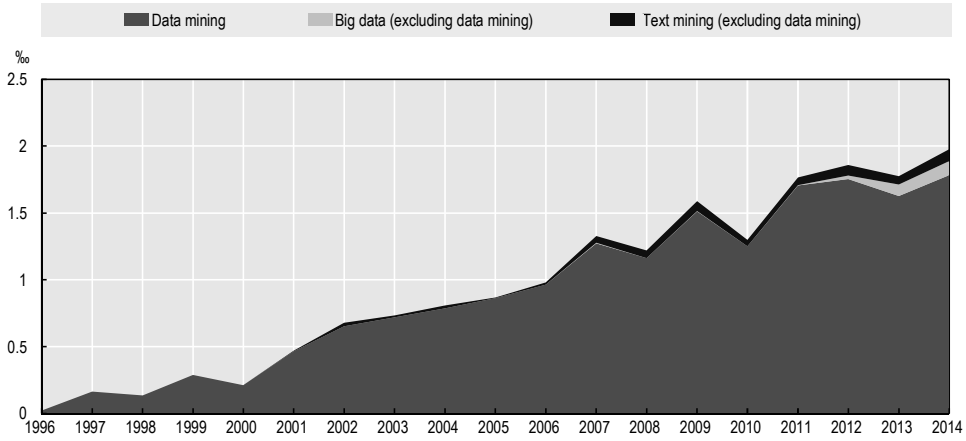
Box 1.3. **The role of an open scientific community for algorithmic transparency**

Increasing transparency related to the functioning of data analytics can be challenging as it may in some cases put at risk proprietary intellectual property rights (IPRs) including trade secrets, which some businesses would consider the "secret sauce" of their business operations (OECD, 2015b). Open scientific communities can play a key role for enhancing algorithmic transparency, while preserving the IPRs of data controllers and increasing awareness about the potentials and risks of data analytics (see Chapter 7 of this volume for further discussions on open science). Many innovative uses of data analytics are disclosed and discussed in conferences and/or scientific papers. Within the last 10 years between 2004 and 2014 scientific articles on data analytics (and related terms) have grown by 9% a year on average (Figure 1.9).

Box 1.3. **The role of an open scientific community for algorithmic transparency (cont.)**

Figure 1.9. **Data analytics related articles in the Science Direct repository, 1995-2014**

Per thousand articles available



Source: OECD, 2014c, *Measuring the Digital Economy: A New Perspective*, based on ScienceDirect repository, www.sciencedirect.com, July 2014.

Two cases are highlighted here as illustrative for the potential of open scientific communities:

- Target's use of data and analytics to predict pregnancy (mentioned above) – In this case, the work was presented in 2010 by a Target statistician at the Predictive Analytics World (PAW) Conference under the title "How Target Gets the Most out of Its Guest Data to Improve Marketing ROI" (Pole, 2010). The author's presentation remained unnoticed by the public for two years, until Duhigg (2012) and later Hill (2012) discussed Target's practice.

- Facebook's experiment on massive-scale contagion conducted by Facebook in 2012 – This experiment was disclosed in Kramer et al. (2014) and based on the manipulation of content presented to more than 689 000 Facebook users. The result showed that "emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks" (Kramer et al., 2014). In the case of Facebook's experiment, it was the paper (Kramer et al., 2014) that revealed the details of the experiment and that was immediately discussed by the public after being published and made openly available through the *Proceedings of the National Academy of Sciences* (PNAS).

Skills and employment in a data-driven economy – The focus of Chapter 6

DDI is disruptive and may induce the "creative destruction" of established businesses and markets. Evidence suggests that creative destruction is an essential engine of long-term economic growth in market economies. In particular in the current context of weak global recovery, business and policy leaders need "to take advantage of the process of 'creative destruction' to accelerate structural shifts towards a stronger and more sustainable economic future" (Guellec and Wunsch-Vincent, 2009). However, managing structural change is challenging, as the pursuit of (short-term) profit may result in

reluctance to change – see what Christensen (1997) refers to as the "innovator's dilemma". In addition, too many businesses (and their employees) may have a weak capacity to change and realise the potential of DDI as suggested by the still low level of ICT adoption across countries (see Figure 1.8).

Structural change in labour markets

DDI may further increase pressure for structural change in labour markets, since it enables the automation of an increasing number of cognitive and manual tasks.[35] This includes the use of data analytics for a wider range of intellectually demanding tasks, such as diagnosis of diseases based on analysis of complex information. It also includes use of a new generation of autonomous machines and robots that are no longer restricted to very precisely defined environments, and that can be deployed and redeployed at much faster rates compared to current generation robots.

Many observers see a high risk that "smart" applications will further broaden employment polarisation, at least in the short run. The effects could be that more middle income jobs may be negatively affected – jobs largely held by the segment of the population that "glues" our societies together. Furthermore, DDI will also affect manufacturing, and in fact could reduce the number of blue collar jobs needed.

Certain categories of jobs, however, are less likely to be susceptible to computerisation and (data-driven) automation. These include jobs that involve solving unstructured problems, including problems that lack rules-based solutions; working with new information, including making sense of new data and information for the purpose of problem solving or decision making, or to influence the decisions of others; and non-routine manual tasks, carrying out physical tasks that cannot be well described via rules because they require optical recognition and fine muscle control that continue to prove difficult for robots to perform. While solving unstructured problems and working with new information will be particularly important for high-end jobs, carrying out non-routine manual tasks will become more and more important for low-paying jobs.

The growing importance of data specialist skills and employment

At the same time, the increasing use of data and analytics across the economy has driven demand for new types of skills and jobs, most prominently involving data specialisation. All these data-specialist professions have one common denominator: working with data constitutes a main part of the job. However, there is currently a relatively low availability of critical data specialist skills and competence required for DDI, and this may prove not only a barrier to adoption of DDI, but also a missed opportunity for job creation: some have suggested that the demand for data specialist skills exceeds the supply on the labour market. An Economist Intelligence Unit (2012) survey, for instance, shows that "shortage of skilled people to analyse the data properly" is indicated as the second biggest impediment to make use of data analytics. There is especially a need for people that possess the skills needed for extracting insights from data.

Employment opportunities will remain for people with the right mix of skills and competencies. The most data-intensive industries employing the highest share of data specialists are still the ICT service industries, and in particular i) IT and other information service industries, ii) insurance and finance, iii) science and research and development, and iv) advertising and market research. But v) the public, and (vi) health care sector are identified as promising areas for new data specialist jobs.

Policy considerations for smoothing structural change

DDI may further increase unemployment and inequality through skill-biased technological change, if not addressed by policy measures – including via social and tax policies. This suggests the need for a smart "double strategy" that promotes continuous education, training and skills development, while addressing the risks of worsening inequality in earnings in labour markets. That need is especially acute given the current weak global recovery and lingering high unemployment in major advanced economies. In the context of DDI, inequality could become a major issue if access to urgently needed high-quality education to take advantage of the job creation opportunities ahead is limited to a few.

Education systems should support a broader interdisciplinary understanding of multiple complex subjects but also deeper insights into some domain-specific issues. Soft skills such as creativity, problem solving and communication skills are key for ensuring employment in a data-driven economy; skills involving fine muscle control will also become a key competitive advantage of humans over machines. These skills, if cultivated with the support of education systems and accompanied by political attention and good co-operative global governance, may lessen concerns related to technological unemployment. This will be more the case, if individuals can enhance and complement their talents to use technology to "dance" with the machines instead of "racing" against them.

The policy considerations discussed in the first six chapters presented above apply across all sectors and application domains. There are in addition domain-specific issues that policy makers need to address, in particular when promoting DDI in specific sectors. As highlighted above, the "low hanging fruit" from the adoption of DDI are in research and education, health care, as well as the public sector. The remaining four chapters presented below focus on these key areas, and highlight specific opportunities and challenges that decision and policy makers need to consider in more detail:

1. Promoting data-driven scientific research – The focus of Chapter 7
2. The evolution of health care in a data-rich environment – The focus of Chapter 8
3. Cities as hubs for data-driven innovation – The focus of Chapter 9
4. Governments leading by example with public sector data – The focus of Chapter 10.

Promoting data-driven scientific research – The focus of Chapter 7

Data analytics now makes it possible to collect, generate, access, use and reuse research and scientific material (articles and data sets but also images and digital lab records) at no or extremely low marginal cost. As a result, the speed at which knowledge can be transferred among researchers and across scientific fields can be increased, opening up new ways of collaborating and new research domains. Different scientific domains are becoming increasingly interconnected: data generated in one field of research may nowadays be treated with models and techniques traditionally belonging to other fields of research. The term often used to describe this transformation of science into a more open and data-driven enterprise is *open science*.

Impacts of open access to scientific data

Data generated at the global level that relate to issues of global concern, such as the environment and climate change or the ageing population and health, may be more

powerfully exploited if properly interconnected and used by large networks of scientists and researchers world-wide. This can avoid duplication of effort and enhance co-ordination in science and research. Furthermore, open scientific data have the potential to strengthen relations between the scientific community and society. Scientists now have a broader range of mechanisms (e.g. through social networks, personal scientific blogs, videos, interviews and discussion forums) to communicate with citizens. And these new scientific communication mechanisms can help build public trust in science.

Collaborative efforts in science and research can thus reach beyond the research community to increasingly involve citizens and "amateur researchers" at different stages of scientific processes, from data collection to solving more complex scientific problems. The involvement of non-professional scientific communities in science and research efforts is often referred to as *citizen science* and it comes with several benefits; for example, it allows the development of a more democratic environment in science by engaging amateurs as well as professionals in research and scientific efforts.

There are other examples of crowdsourcing for technical skills that can solve scientific problems, such as online platforms where solutions to scientific problems are requested from the public. Private companies and research teams publish unsolved problems related to specific data sets (also published on the platform), and data scientists from all over the world compete to find the best solutions and highest-performing algorithms for prize money. The approach relies on the fact that there are countless strategies to solve the problem, each with a different computational efficiency.

The involvement of citizens in scientific projects tends to have an educational value, both implicit and explicit. While in the majority of projects the informal learning aspect of adult citizens is addressed, schools are increasingly considered an important target for the introduction and promotion of citizen science. Teachers play a significant role in facilitating the deployment of experiments and transmitting the socio-scientific values of their contributions to the young audience. In fact, a number of countries are investing in the educational skills building necessary for data analytics, as these skills are currently lacking.

Greater access to scientific inputs and outputs can improve the effectiveness and productivity of the scientific and research system, by reducing duplication costs in collecting, creating, transferring and reusing data and scientific material; by allowing more research from the same data; by multiplying opportunities for domestic and global participation in the research process; and by ensuring more possibilities for testing and validating scientific results. With unrestricted access to publications and data, firms and individuals may use and reuse scientific outputs to produce new products and services. Developing countries in particular may benefit from open access to scientific material.

Challenges to open data and data sharing in science and research

However, several barriers to data sharing still remain. Some are of a technical nature, such as issues related to storage, the technical infrastructure to allow data sharing, interoperability and standards. Other types of barriers are related to the lack of an open data culture or the disincentives that researchers and scientists face with respect to the disclosure and sharing of data sets, especially relative to research at the pre-publication stage. This raises the question of the "optimal" level of openness to boost research and innovation without discouraging data collection from individual researchers.

Additional challenges relate to the definition of ownership of the data itself. Barriers to legal, cultural, language and proprietary rights of access hinder cross-national collaboration and international data exploitation, especially in the social sciences. There are issues with regard to propriety databases that could impede open research data efforts in academia as well. There is thus tension between open research data and IPRs, and a balance must be struck between efforts to promote open data in science and efforts to promote commercialisation of public research, especially in the case of public-private partnerships involving companies. The tension can however be lessened by policies that clarify IP ownership and promote non-exclusive licensing possibilities, as well as by greater IP awareness among researchers.

Another legal issue that comes into play in the context of open scientific data is privacy and personal data protection. Data gathered in the course of research often contain personal information (e.g. medical records), and so opening such data has to respect the rights of data subjects (Lane et al., 2014). This does not mean that the data cannot be opened, but it does calls for implementing effective protective procedures.

Data collection, curation and sharing vary by scientific discipline; some fields have been traditionally more data-intensive than others. Researchers belonging to scientific disciplines not involving large-scale experiments managed by teams of hundreds of researchers, notably in the social sciences and humanities, traditionally collect and built their own data sets, in some cases manually or by developing surveys and questionnaires. That makes this kind of the data set more tied to the individual researcher, and therefore less easily ready to be shared without proper curation, cleaning and metadata compilation. Scientists and researchers do not have necessarily the incentives or the skills to perform those tasks, since proper curation and dissemination of data sets are costly and time-consuming. Also, they traditionally compete to be first to publish scientific results, and may not see the benefits of disclosing information on the data they want to use to produce as yet unpublished research outcomes.

A possible solution to the above-mentioned disincentives is data citation. Researchers wishing to be acknowledged for their work could release data sets through mechanisms similar to the one already in place for citations of academic articles. Data citation is not, however, necessarily a standardised or widely accepted concept in the academic community.

The evolution of health care in a data-rich environment – The focus of Chapter 8

The health sector is a knowledge-intensive industry: it depends on data and analytics to improve therapies and practices. There has been tremendous growth in the range of information being collected, including clinical, genetic, behavioural, environmental, financial and operational data. Every day, health care professionals, biomedical researchers and patients produce huge amounts of data of great value from an array of devices. At the same time, the potential to process and analyse these emerging multiple streams of big data and to link and integrate them is growing.

Drivers of growth of digitised health data

Five principal factors drive the increased collection and use of large-scale data in the health sector. One is demographic change and a 20-year shift in the burden of disease, from infectious conditions to long-term non-communicable diseases (NCDs) brought on by lifestyle choices and environments. A second factor is that fiscal pressures have led to

a need for greater efficiencies. Continuing pressure to find ways to make systems more productive has moved the focus from cost containment to performance-based governance. To evaluate health sector performance, managers and governments will need timely and accurate information about the prices and volumes of services provided and the health outcomes produced, at levels sufficiently detailed to take corrective policy action.

The third has to do with the role of the patients themselves in the care process, which has taken on much greater importance in recent years. Patients' taking command of and managing their health will especially aid in the management of chronic diseases. The fourth driver is the need for co-operation to tackle global public health challenges such as infectious diseases, and improve early detection and warning of emerging health threats and events. Complementing the traditional case-based and syndromic surveillance systems, monitoring of unstructured events – through news and Internet media, web searches, etc. – has been a significant component of public health early warning and response over the past decade.

But the fifth driver of health data use is possibly most important: the sheer volume, velocity and variety of health data available. Many health care systems are rapidly digitising immense amounts of data and using them for a wide range of activities, including preventive care, e.g. early detection; field data to support emergency and urgent care; coaching, rehabilitation and maintenance; context-sensitive intervention, e.g. reminders; epidemiological assessments; post-market surveillance and analysis; health care quality and performance monitoring.

The increasing use of electronic medical records promotes patients' participation – in their own care, in self-management of health conditions, and in informed decision making. Patients' interest in their diagnostic test results and medical records, in their options for care, in the quality of providers, and in scheduling visits on line will keep growing. Over the past decade multiple studies have documented the value of electronic personal records (EPRs) in supporting greater patient-centred services. Patients and practitioners are also increasingly interested in devices, tools and computer applications that assist in monitoring and improving health and well-being. They recognise that these can help patients live longer in their own homes rather than in considerably more expensive hospital or nursing home facilities; and encourage personal responsibility for healthier lifestyles.

Towards smarter models of care

Any systematic effort to address today's health and wellness challenges will also require data to support new and "smarter" models of care. That will require enhanced capacity for the sharing, processing and analysis of health and behavioural data to support patient-centric care, and a more efficient clinical research enterprise for improved prevention and better disease management. Today's care is reactive, episodic and focused on disease. The new health care will need to be proactive, preventive, and focused on quality of life and well-being.

The ubiquitous care model, for example, is based on the utilisation of smart sensing and biometric devices for real-time monitoring, analysis and transmission of health data. The information can then be accessed by health care providers for informed diagnosis, clinical decisions regarding treatments, and evaluation of outcomes. It can also be viewed and acted upon by patients for both education and prevention.

Mobile health (mHealth) offers a wide range of smart modalities by which patients can interact with health professionals, or with systems that can provide helpful real-time feedback along the care continuum, from prevention to diagnosis, treatment and monitoring. mHealth is of particular value in managing health conditions where continuous interaction is important, such as diabetes and cardiac disease. The devices utilised include mobile phones, tablets, global positioning system (GPS) devices, mobile tele-care devices, and mobile patient monitoring devices.

Crowdsourcing is emerging as a means of allowing science to be conducted at scales of magnitude greater than before. It involves capitalising on the Internet and large groups of people, particularly via online Web 2.0 communities, to harvest "collective intelligence" and accomplish tasks that might have traditionally been given to small research groups. Crowdsourcing can help process data quickly, on unprecedented scales, and with better quality control than any individual or small research group can attain. Crowdsourcing therefore has cost and speed benefits, although careful attention must be paid to policy regarding in particular privacy, security, and data stewardship.

Critical success factors and policy priorities

In fact, a number of challenges must be overcome before the benefits from DDI in the health sector can be reaped. One of these is that electronic health records – EHRs – are being collected in health care systems that are often fragmented, with points of care functioning as silos. Questions of privacy also have to be addressed, and skill building will be needed to analyse the voluminous health data sets. Standards and interoperability are other central issues that must be addressed: while health care organisations have access to an ever-increasing number of information technology products, many of these systems cannot "talk" to each other, and if the systems cannot communicate, big data will not meet its potential in the health care system. Attention is also needed to ensure that individuals who wish to restrict or withdraw their data from their contribution to research and statistics can reasonably do so.

Additional critical success factors for governments to realise value from investments in health data are strategic planning; ensuring legislative and regulatory requirements that support planning; engaging all stakeholders in planning and governance; promoting global co-operation; setting standards for data governance; and providing financial stimulus toward data development and use.

Cities as hubs for data-driven innovation – The focus of Chapter 9

A large share of the 65 million sensors estimated to be deployed (e.g. in security, health care, the environment, transport and utilities) are today embedded in urban infrastructures, facilities and environments (MGI, 2011). With around three-quarters of the OECD area population expected to be living in urban areas by 2022, cities will host at least 10 billion out of the 14 billion devices estimated to be in use in member countries by then (OECD, 2010; OECD, 2012). This makes cities a potential hot spot for DDI.

The urban data ecosystem

The data produced and collected in cities can be divided into three categories. There are data on flows; sensors embedded in urban infrastructures increasingly allow the digitisation and datafication of flows of resources, products, people and information across cities. Data on states of urban spaces and environments, subject to constant natural and manmade changes – the density of people or things (e.g. vehicles), air temperature

and quality, light and sound levels, etc. – are monitored by in situ sensors. Finally, data can relate to activities – transaction, consumption and communication patterns that include people's personal and professional activities, communication and interactions; interaction between people and their environment; and interactions among components of their environments, such as communicating or autonomous machines and devices.

Many actors are involved in data collection and use in cities. Key among them are citizens and consumers; innovators and entrepreneurs; governments and utilities; data brokers and platforms; and infrastructure and system operators. Each of them is in principle connected to all the others, through a digital layer and in multiple possible combinations. The extent to which data can be exchanged among these actors and across systems in cities, as well as the extent to which they can easily be reused for different purposes, determines their potential for DDI.

Opportunities for data-driven innovation in cities

Much of the data on flows and states in cities, and some of the data on activities, can be used to increase the efficiency of urban systems and promote their integration. The availability of historical and real-time data on flows in transport, energy, water and waste systems enables analysis at unprecedented depth and granularity, as well as targeted interventions in and precise management of urban systems. So far, the most promising effects of information and communication technologies (ICTs) and data use in cities can be found in transport and electricity, two systems that share an important lever for data-driven improvements: the direct match of demand and supply, based on fuller and often real-time information.

Synergies can be reaped through integration of these systems. Understanding urban infrastructures and sectors as systems, a city can be considered a "system of systems", within which ICTs and the digitisation of urban flows are creating the potential for deep integration (CEPS, 2014). The Internet of Things will continue to multiply the systems, machines, devices and services connected via electricity grids and information systems – such as solar cells on roofs, detailed weather forecasts, home heating systems and air conditioning, supermarket stocks, etc.

Over the past years, innovative start-ups have penetrated established urban sectors with data-driven mobile apps and online platforms. Known under the label "sharing economy", new business models are using real-time and geo-locational data on online platforms and mobile apps that allow commercial "sharing" (renting) of cars, rides and bikes as well as vacant homes, offices and shops in cities. On the supply side, car owners can rent their car if they are not using it, sell seats on trips they are taking anyway, or work as private drivers when time permits; real estate owners can rent out vacant living, office or commercial space for short periods. On the demand side, urbanites get more and cheaper mobility options, and travellers a larger and cheaper choice of accommodations; freelancers and the self-employed gain flexible access to office and commercial space.

City administrations are also increasingly using urban (crowdsourced) data to gain fine-grained real-time information on aspects such as public service delivery, system performance and infrastructure conditions. Mobile apps now allow citizens to report on stray garbage, potholes, broken lamps and the like via their smartphone, directly to city hall. Online, crowdsourced, real-time and geo-locational data can also play an important role in disaster management in cities. The UN Global Pulse project, for instance, uses real-time analytics to turn unstructured online information into actionable information for decision makers to improve resilience.

Greater data availability and more powerful computing are bringing urban modelling back into the spotlight of urban planning, and have the potential to significantly improve the forecasting of societal demand. Geo-referenced data collected via (e.g.) crowdsourcing, remote sensing and social networking – combined with new computational power, including cloud computing – offer fresh possibilities, notably as applied to integrated land use and transport planning (Nordregio, 2014). Data analysis and modelling of societal demand for urban infrastructures and services have the potential to significantly improve resource allocation and investment decisions in urban areas.

Challenges and policy priorities

There are opportunities to be seized in spurring DDI in cities, but there are also challenges. Cities need to build the requisite capacity and skills for collecting, storing and analysing data in a depth and at a scale that are unprecedented, in addition to acquiring the infrastructure and computing power needed to store and process all the data. Sensitive questions need to be addressed when it comes to the type of data cities should collect in the first place and what they should publish thereafter. An important condition for advancing integration of urban systems and system-to-system communication is interoperability across different systems and components at different levels.

Cross-sector data sharing is likely to pose challenges. Data collected in different sectors tend to be stored in different formats, and few incentives exist for harmonising them. Without open standards, data sharing may be limited by and locked into proprietary formats. Linked issues are privacy protection, overcoming silo structures in administrations, and improving co-operation among jurisdictions and levels of government.

While increasing system integration in cities can yield benefits, it also creates new risks. The more ICT, energy, transport and other critical urban infrastructures and systems are interconnected, the more a city as a system-of-systems will become vulnerable to both internal and external threats, ranging from technical failures to cyber attacks and natural disasters. That vulnerability calls for a digital security risk management framework to reduce risk to an acceptable level in light of the expected benefits, through security and preparedness measures that fully support the economic and social objectives at stake.

Governments leading by example with public sector data – The focus of Chapter 10

The public sector is one of the economy's most data-intensive sectors. Its importance as an actor in the data ecosystem is twofold: as a key user of data and analytics, and as a key producer of data that can be reused for new or enhanced products and processes across the economy. The idea behind open access to public sector data is that value can be derived through the reuse of that data by any user from within or outside the public sector. Governments can therefore promote DDI by leading by example in their use and supply of public sector data.

The potential of public sector data

For government – The use of open government data (OGD) by government agencies can lead to efficiency improvements in the public sector. It can, for example, help bring down silos and foster collaboration across and within public agencies and departments. Furthermore, the increasing amount of data made available in formats that enable reuse and linkage is supporting the expansion of data analytics in the public sector; here too,

there is great potential for value creation. Predictive data analytics can, for example, facilitate identification of emerging governmental and societal needs. Use of this data by the public sector can also make for better decisions, inform policies, support the development of data-driven processes and services, and deliver more innovative services. There are also, of course, considerable risks in governments' use of data analytics, in particular with regard to the privacy of citizens.

For citizens – Open government advocates believe that OGD can be a powerful force for public accountability, by making existing information easier to process, combine and analyse. OGD can then promote greater transparency, and allow a new level of public scrutiny that can increase public accountability. E-participation also aims at enhancing citizens' engagement in public life, e.g. in lawmaking, policy making and service design and delivery. Citizens become not just passive consumers of public sector content and services but also active contributors and designers in their own right, empowered to make more informed decisions that can enhance the quality of their lives.

For the private sector – First of all, granting the private sector better access to public sector information (including public sector data) can increase efficiency, effectiveness and innovation in public service delivery. The strategy is to provide innovators from outside governments with the opportunity to develop modular services that are more agile and targeted to citizens' needs than those developed in-house by governments. Secondly, as the importance of data in the development of new services, products and markets has increased dramatically, open access to public sector data can stimulate innovation in the course of that development as promoted by the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* (PSI).

The OECD market for PSI was estimated to be around USD 97 billion in 2008, and could have grown to around USD 111 billion by 2010. Aggregate OECD economic impacts of PSI-related applications and use were estimated to be around USD 500 billion, and there could be close to USD 200 billion of additional gains if barriers to use are removed, skills enhanced and the data infrastructure improved. There is cross-country evidence that significant firm-level benefits are to be had from free or marginal cost pricing, with small and medium-sized enterprises (SMEs) benefiting most from less expensive data.

Key challenges in implementing open data and PSI strategies

The lack of procedures and standards on how to deal with open data in governments can compromise the quality of the data and eventually the output of OGD and PSI initiatives. Public sector data often are not harmonised, making it difficult from the user perspective to know which data are valid or should be trusted. Critical to access is knowing the source of what one is searching for, and in many instances where to start searching is a challenge. Accessibility can also be limited if data cannot be reused, and data transparency may be hindered if data are not simple to access or reuse due to their format. Interoperability is equally a priority concern for policy makers tasked with implementing OGD or PSI strategies.

The economic climate undoubtedly plays a role. At a time of budget pressures and cuts in government expenditures, it is important to articulate clearly the advantages of opening up public sector data for wider use and, where necessary, to compensate the providers of public sector data for any initial extra funding necessary to open up and digitise the data. Consequently, great emphasis is now being placed on devising more solid methodologies to assess the impact of open access.

Finally, there are organisational, cultural and legal challenges. Having a consistent legal framework in place is critical; fragmented and diverse legislation concerning privacy, reuse of data and related fees can create confusion for end users. And legislation, IT platforms and codes need to be matched by a culture within the public service that supports a presumption to publish, release and share data. Raising awareness of civil servants, citizens, civil society organisations and the private sector with regard to their rights is important for society as a whole to fully capture the benefits of public sector data.

1.3. Common key challenges and policy considerations

DDI is disruptive and comes with major economic and societal challenges and risks to be addressed across all application areas and sectors. Some of the challenges are the result of serious tensions between opposing private and social (collective) interests. Addressing these tensions is complex and cannot be undertaken in silos; these require governments to invite the democratic participation of all citizens – in addition to stakeholders including civil society, the technical Internet community and business groups – in order to be resolved. This also calls for a whole-of-government strategy to promote DDI, as countries such as Australia,[36] Japan,[37] and the United Kingdom[38] as well as the European Commission (EC)[39] have developed or are envisioning.[40]

Two sets of challenges (tensions) need to be addressed by policy makers in order to maximise the benefits of DDI, and mitigate the associated economic and societal risks. A set of key policy issues discussed across the chapters of this volume are related to the need to i) promote "openness" in the global data ecosystem and thus the free flow of data across nations, sectors, and organisations, and at the same time ii) address legitimate considerations of individuals' and organisations' opposing interests (including in particular their interests in the protection of their privacy and their intellectual property rights).[41] Another set of policy issues aims at iii) activating the enablers of DDI, and at the same time ii) addressing the effects of the "creative destruction" induced by DDI, in particular with a focus on small and medium enterprises (SMEs) and on labour markets.

These two sets of tensions may at first appear unrelated. However, a closer look at the policy issues discussed across the chapters suggests that a move towards more "openness" may further the disruptive effects of DDI and thus lead to more "creative destruction". That said, there is no one-size-fits-all optimal level of "openness"; instead the optimal level strongly depends on the domain and the cultural environment in question. Furthermore, in addressing these tensions, policy makers should be aware of the "path-dependency" of current actions (and inaction) that could limit future choice. History of the diffusion of new standards provides examples of path-dependency where early adoption may prevent a more efficient standard at a later stage.[42] In the case of DDI, path-dependency is not only related to the adoption of standards, which play a key role for "openness" in the data ecosystem. The interaction of individuals with data analytics can shape their preferences (including for privacy), and set society on a path that could become impossible to change in the future. This calls for a careful assessment of current policy actions (and inaction) to maximise the long-term benefits of DDI.

Finally, policy makers should acknowledge that DDI may favour concentration and greater information asymmetry and thus shifts in power. This may lead to a new digital (data) divide that could undermine social cohesion and economic resilience. As discussed above, the economic value of DDI is reaped when better insights (knowledge) can be extracted from data. With this knowledge come better insights and more capacity to

influence and control. Where the agglomeration of data leads to greater information asymmetry, power could shift away: (i) from individuals to organisations (incl. consumer to business, and citizen to governments); from traditional businesses to data-driven businesses given potential risks of market concentration and dominance; (iii) from governments to data-driven businesses, where businesses can gain much more knowledge about citizens (and politicians) than governments can; and (iv) from lagging economies to data-driven economies.

Overall, countries will be able to maximise the benefits of DDI, if they can connect to the global data ecosystem (Chapter 2), leverage the enabling factors of DDI (Chapters 3) and promote investments in data as infrastructure (Chapter 4), and address the various key policy challenges (Chapters 5-6), including the domain-specific ones (Chapters 7-10). Given all of this, governments have an important role to play in promoting DDI and mitigating the associated risks.

Annex – Highlights of the 2014 Global Forum on the Knowledge Economy

On the occasion of the 2014 OECD Ministerial Council Meeting, under the Chairmanship of Japan on the 50th anniversary of its accession to the OECD, Ministers affirmed the importance of knowledge-based capital to provide new sources of growth in the face of long-term challenges, such as ageing and environmental degradation, and that the OECD's work on the digital economy is important.

The 4th Global Forum on the Knowledge Economy (GFKE) held in Tokyo, Japan, on 2-3 October 2014, focused on data, one example of knowledge-based capital. Policy makers, business, civil society and other stakeholders from OECD Member and Partner (i.e. non-member) economies participated in active discussions on data-driven innovation for a resilient society.

Throughout the entire forum, participants acknowledged the high value of big data in spurring economic growth or solving various social challenges, and discussed policy options to promote the use of big data that will inform the discussion at future OECD meetings. Highlights of the discussions include:

1. *Illustrating the economic benefits* – Participants discussed the positive economic impacts of big data across industries, and in particular manufacturing, and emphasised that data-driven innovation is likely to promote economic growth in both OECD member and non-member economies, directly or through spillover effects. Participants mentioned the value of optimising existing services and of analytics for decision making. Participants discussed the global dimensions of data-driven innovation, including the importance of cross-border data flows for trade, as well as the need to understand and address the implications of data-driven innovation for jobs.
2. *Addressing complex societal challenges* – Participants recognised the potential of big data analysis for disaster response (for example, based on the ex post analysis of the Great East Japan Earthquake), but also more generally for improving quality of life. They underlined the need for government leadership, awareness and collaboration among all actors in the adoption and implementation of disaster risk management approaches to enhance human security. As an example, it was shown how big data can be used to relieve traffic congestion and improve construction standards.
3. *Leveraging data-driven innovation in ageing societies* – Participants recognised the opportunities that data-driven innovation presents for ageing societies, but agreed that most of the potential for value creation is still unclaimed. They discussed the need to overcome data silos and create the appropriate conditions for broader data access, linkage and integration. It was recognised that local data on vulnerable elderly populations are necessary for central government actions and disaster planning. Defining minimum standards for data was considered essential, as well as interoperability. An important idea put forward was the need to create conditions for a risk-based approach to protect data. Finally, participants concluded that there is a need to strengthen the capacity to analyse data, build expertise, and increase business opportunities.

4. *Promoting skills for the data-driven economy* – Participants were aware of the gap between the demand and supply of data scientists, and the need for skills development and education. Potential displacement effects were highlighted in particular with regard to certain middle income, white collar jobs as well as the need to address the resulting inequality implications. Problem solving and entrepreneurial competences building on human creativity and intuition, in combination with data analysis and software engineering skills, were highlighted as critical as well as basic ICT literacy. Participants recognised the importance of lifelong learning as a means to fill the potential employment gap.
5. *Building trust in the data-driven economy* – Participants recognised that the trust of individuals is crucial and that big data users should respect fundamental values. They underlined the importance of risk-based approaches to the collection and use of personal data. Algorithmic transparency raises complex issues, but providing information on key elements informing decisions is important to avoid discrimination. Other key issues discussed included security, ethics, privacy-enhancing technologies and better metrics. The impact of data concentration on privacy, but also on competition, transparency and accountability, was considered worthy of further examination.
6. *Encouraging open data across society* – Participants underlined the necessity of promoting open data so as to make it possible to use public data to create new services and effective administrative procedures. Public value depends on data use. In response, governments' role is evolving from the direct provision of data and regulation. It also now encompasses the creation of enabling conditions supporting communities of providers and users, building trust, enforcing principles of non-discrimination for public entities, civil society and the private sector to improve open data sharing and use.
7. *Policy conclusions* – Governments and stakeholders need to develop a coherent policy approach to harness the economic benefits of data-driven innovation. They need to assess the context for data collection, analysis and use to ensure that data-driven innovation serves societal values in an ethical and equitable manner.

Notes

1 As a point of reference, one exabyte corresponds to one billion gigabytes and is equivalent, for example, to around 50 000 years of DVD-quality video (see http://exabyte.bris.ac.uk/, accessed 15 May 2015).

2 Estimates are provided by IBM (see www-01.ibm.com/software/data/bigdata/what-is-big-data.html, accessed 15 May 2015).

3 It is estimated that 90% of all this data were created in the past few years (ScienceDaily, 2013; Wall, 2014).

4 These include Ministers from and Representatives of Australia, Austria, Belgium, Canada, Chile, Colombia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom, the United States and the European Union.

5 As stated in the OECD (2014a) Ministerial Council Statement, "Rising inequality endangers social cohesion and weakens social resilience, thereby hampering economic resilience. A key challenge is to achieve inclusive growth by providing social protection and empowerment to people, which can strengthen human security. Appropriate flexibility and security in labour markets and relevant education and skill programmes can facilitate greater inclusion and participation of under-represented groups. We welcome OECD initiatives targeting these groups, including on gender equality, youth employment, ageing society and the integration of migrants. We also recognise that regional and urban policies can play a key role in empowering people and building resilience at all levels of our economies and societies".

6 The outcomes of the first phase of the OECD horizontal project on *New Sources of Growth: Knowledge-Based Capital* (KBC1, see OECD, 2013a) were discussed at the conference on "Growth, Innovation and Competitiveness: Maximising The Benefits Of Knowledge-Based Capital" on 13-14 February 2013, and the final conclusions were presented to ministers at the 2013 OECD Ministerial Council Meeting (MCM) (see http://oe.cd/kbcconference).

7 Calculated based on annual balance sheet data as follows: $(p - d) / a$, where p = the total gross value for property, plant, and equipment; d = total accumulated depreciation; and a = total assets.

8 Brynjolfsson et al. (2008) highlighted how information technology (IT) had "enabled firms to more rapidly replicate improved business processes throughout an organization, thereby not only increasing productivity but also market share and market value". Internet firms, however, are not only replicating business processes throughout their organisations, but also increasingly relying on automated business processes that are empowered by software and in particular data analytics.

9 The growth of storage-related technologies and services can be explained by the fact that many top ICT companies are trying to strengthen their market position through the development of new "big data" branded products, many of which are based on open source storage management solutions initially developed by Internet firms such as Hadoop, a major big data technology (see Chapter 2 of this volume).

10 The market value of (traditional) relational database management systems alone was worth more than USD 21 billion in 2011, having grown on average by 8% a year since 2002 according to some estimates (OECD, 2013b).

11 *Social Genome* builds on public data from the web (including social media data) as well as Walmart's proprietary data, such as its customer purchasing and contact data.

12 For example to analyse and predict potentially vulnerable components; the resulting analysis is further used to optimise product design and production control.

13 Similar services are observed in the energy production equipment sector, where M2M and sensor data are used to optimise contingencies in complex project planning activities for instance (Chick, Netessine and Huchzermeier, 2014).

14 Some of the data and analysis results are presented to farmers via the MyJohnDeere.com platform (and its related apps), to empower farmers to optimise the selection of crops and the time and place for planting and ploughing them (Big Data Startups, 2013).

15 The study is based on a survey by Bakhshi and Mateos-Garcia (2012), but extended by "matching survey responses about data activities with historical performance measures taken from respondents' company accounts, and by conducting an econometric analysis of the link between business performance and data activity while controlling for other characteristics of the business". The analysis shows that, other things being equal, a one-standard deviation greater use of online data is associated with an 8% higher-level of total factor productivity (TFP). Firms in the top quartile of online data use are 13% more productive than those in the bottom quartile. The study furthermore shows that "use of data analysis" and "reporting of data-driven insights" have the strongest link with productivity growth, "whereas amassing data has little or no effect on its own" (Bakhshi, Bravo-Biosca and Mateos-Garcia, 2014). Another study by Barua, Mani and Mukherjee (2013) suggests that improving the quality of and access to data by 10%, by presenting data more concisely and consistently across platforms and allowing it to be more easily manipulated, would increase labour productivity by 14% on average, but with significant cross-industry variations.

16 The estimated output elasticity of 3% resulted after controlling for firms' adoption of data-driven decision making. The OLS (ordinary least squares) estimate on the Hadoop measure indicated an output of 10%, which Tambe (2014) attributed to other omitted variable bias, including firms' adoption of data-driven decision making.

17 See also a survey by the *Economist Intelligence Unit* (2012) of business executives, according to which expectations are that the use of "big data" could improve organisational performance by 25% and by more than 40% over the next three years. The use of data analytics by businesses depends primarily on the type of data sets used. Business activity data and point-of-sale data are more frequently subject to data analytics, whereas online data including social media data and clickstream data are less frequently used among firms across the economy. According to the survey by the

Economist Intelligence Unit (2012), of more than 600 business executives around the world, two-thirds "say that the collection and analysis of data underpins their firm's business strategy and day-to-day decision-making". The respondents considered in particular "business activity data" as the most valuable data sets and in the case of the consumer goods and retail sector, "point-of-sale data" as well.

18 For instance, it is unclear whether those firms adopting DDI became more productive due to DDI-related investments or whether they were more productive in the first place. Furthermore, these studies rarely control for the possibility that some firms may have eventually seen a reduction in their productivity due to DDI, and as a result may have discontinued their investments in DDI.

19 As Mandel (2012) highlights: "[…] economic and regulatory policymakers around the world are not getting the data they need to understand the importance of data for the economy. Consider this: The Bureau of Economic Analysis […] will tell you how much Americans increased their consumption of jewelry and watches in 2011, but offers no information about the growing use of mobile apps or online tax preparation programs. Eurostat […] reports how much European businesses invested in buildings and equipment in 2010, but not how much those same businesses spent on consumer or business databases. And the World Trade Organization publishes figures on the flow of clothing from Asia to the United States, but no official agency tracks the very valuable flow of data back and forth across the Pacific".

20 Occupations were identified from the O*NET database, which provides ratings for hundreds of occupations in relation to many different features including working activities and the level and importance of those activities. Working activities considered for identifying potential occupations included: i) "getting information", ii) "processing information", and iii) "analysing data or information", with the level and importance of all three activities above the 75th percentile, and iv) "interacting with computers" at a level and importance below the 75th percentile. In the health care sector, potential occupations included, for instance, registered nurses, physicians and surgeons, and radiologists.

21 These estimated numbers should be taken with a great deal of caution, given that their underlying methodologies and data are not available.

22 Caution should be exercised when interpreting these results, as the methodologies used for these estimates are not available.

23 These initiatives are based on research results providing evidence of a link between real-world events and spikes in the volume of Twitter conversations related to food prices in Indonesia, illustrating the potential value of employing regular social media analysis for early warning and impact monitoring.

24 Crowdsourcing is "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community" (Merriam-Webster, 2014).

25 Tsaboin also creates opportunities for others to innovate around the traffic data collected by making the data accessible to everyone.

26 The pilot of Cignifi's credit scoring platform in Brazil was conducted with data from pre-paid mobile customers located in the northeast of the country, one of the poorest regions, and provided evidence that the risk score calculated can be a significant

discriminator of default risk. There is great demand for such services in regions with very low levels of financial penetration.

27 Some major policy issues related to DDI have not been addressed in depth in this volume, in particular in respect to the complementary effects between data/analytics and the other types of KBC. This includes the role of intellectual property rights (IPR), which are highly relevant and often used as complementary assets to data and analytics, and sometimes therefore as strategic points for controlling DDI-related activities (Chapter 2). It is therefore recommended that this report be read in conjunction with OECD (2015a), which focuses on the second KBC pillar on IPR. The complementary effects in regard to the third KBC pillar, economic competencies, are highlighted in this report in terms of the skills and organisational change needed to realise the potential of DDI (see Chapter 6; see also Squicciarini and Le Mouel, 2012).

28 The data ecosystem relies on other key elements including technologies and actors that are rarely represented in a simplified model of the system. In particular, Figure 2.2 in the next chapter abstracts from the "cytoplasm" that lies between the layers of the data ecosystem and that enables the smooth interoperability of the different actors, their technologies, and services. These include (open) standards, some of which are related to application programming interfaces (APIs). Representations such as Figure 2.2 tend to be strongly biased toward the ICT sector, and do not sufficiently take into account other roles that are key to the functioning of the data ecosystem (e.g. legal consultants to address privacy risks). Ignoring these other actors will lead to systematic underestimation of the full size and impact of the data ecosystem.

29 Anderson (2008) has even gone so far as to challenge the usefulness of models in an age of massive data sets, arguing that with big data, machines can detect complex patterns and relationships that are invisible to researchers. The "data deluge", he concludes, makes the scientific method obsolete, because correlation is enough (Anderson, 2008; Bollier, 2010).

30 This property is at the source of significant spillovers which provide the major theoretical link to total factor productivity growth according to a number of scholars including Corrado et al. (2009).

31 See also OECD, 2015a according to which "[c]oncerns about government access requests – particularly to data entrusted to providers of cloud computing services – predate the revelations by Edward Snowden in 2013 and are not limited to intelligence gathering. But those revelations have brought into sharper focus the need for transparency. Today, Internet and communications businesses are under increasing pressure to be open about the manner in which they address government access requests."

32 A data breach is "a loss, unauthorised access to or disclosure of personal data as a result of a failure of the organisation to effectively safeguard the data" (OECD, 2012). Where the security breach of intellectual property does not involve personal data, the term "unauthorised access" will be used instead.

33 Duhigg (2012) describes the analysis process as follow: "[…] Lots of people buy lotion, but one of Pole's colleagues noticed that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women

loaded up on supplements like calcium, magnesium and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitisers and washcloths, it signals they could be getting close to their delivery date". As data analytics is not perfect, false positives are to be accounted for (see Harford, 2014). Target therefore mixes up its offers with coupons that are not specific to pregnancy (Piatetsky, 2014).

34 At the fourth meeting of the OECD Global Forum on the Knowledge Economy (GFKE) on "Data-driven Innovation for a Resilient Society" held in 2-3 October 2014 in Tokyo, Japan (www.gfke2014.jp/), Electronic Privacy Information Center (EPIC) President, Marc Rotenberg, highlighted the need for "algorithmic transparency", which would make public data processes that impact individuals (see Annex of Chapter 1 of this volume on the highlights of the *2014* GFKE*)*.

35 Several authors including Ford (2009), Cowen (2013), Mayer-Schönberger and Cukier (2013), Frey and Osborne (2013), Levy and Murnane (2013), Brynjolfsson and McAfee (2014), Rifkin (2014) and Elliott (2014), have highlighted the potential negative implications of data-driven automation on wage and income inequalities.

36 Australia's National Digital Economy Strategy (NDES) foresees under its action item 12 the release of its national "Big Data strategy".

37 The national strategy presented in Japan's *Declaration to be the World's Most Advanced IT Nation* highlights the promotion of open and big data.

38 In 2013, the government of the United Kingdom published its "government strategy for how the United Kingdom can be at the forefront of extracting knowledge and value from data" (see www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf, accessed 12 May 2015).

39 In July 2014, the EC outlined its strategy Towards a Thriving Data-driven Economy in its Communication COM(2014) 442 final. The strategy aims at "supporting and accelerating the transition towards a data-driven economy in Europe".

40 Other countries have established, or are about to establish, sector- or domain-specific big data strategies. In 2012 for instance, the United States released its Big Data Research & Development Initiatives, which foresees investments worth USD 200 million in new R&D (see www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf, accessed 14 May 2015) as well as its strategic paper on "Big Data: Seizing Opportunities, Preserving Values" (EOP, 2014; see also PCAST, 2014). In 2012, Korea established its Big Data Master Plan, which promotes step-by-step big data use. This includes in particular the establishment of an infrastructure for big data sharing, and the provision of technical support and expert training.

41 In addressing privacy considerations, for instance, policymakers should seek to preserve the openness of the data ecosystem and the Internet.

42 The difficulties in transitioning from IPv4 standards towards the more efficient IPv6 are well known to policy makers (see OECD, 2014d).

References

Acatech (2013), "Securing the future of German manufacturing industry: Recommendations for implementing the strategic initiative INDUSTRIE 4.0", Final report of the Industrie 4.0 Working Group, April, www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Material_fuer_Sonderseiten/Industrie_4.0/Final_report__Industrie_4.0_accessible.pdf.

Adner, R. (2006), "Match your innovation strategy to your innovation ecosystem", *Harvard Business Review*, April, http://pds12.egloos.com/pds/200811/07/31/R0604Fp2.pdf, accessed 15 June 2015.

Allen, K. (2013), "Global financial crisis hit happiness and trust in governments – OECD", *The Guardian*, 5 November, www.theguardian.com/business/2013/nov/05/global-financial-crisis-happiness-trust-governments-oecd.

Anderson, C. (2012), "The man who makes the future: *Wired* icon Marc Andreessen", *Wired*, 24 April, www.wired.com/2012/04/ff_andreessen/5/, accessed 5 May 2015.

Anderson, C. (2008), "The end of theory: The data deluge makes the scientific method obsolete", Wired Magazine, 23 June, www.wired.com/science/discoveries/magazine/16-07/pb_theory/.

Bakhshi, H., A. Bravo-Biosca, and J. Mateos-Garcia, (2014), "Inside the datavores: Estimating the effect of data and online analytics on firm performance", Nesta, March, www.nesta.org.uk/sites/default/files/inside_the_datavores_technical_report.pdf, accessed 13 May 2015.

Bakhshi, H. and J. Mateos-Garcia (2012), "Rise of the datavores: How UK businesses analyse and use online data", Nesta, November, www.nesta.org.uk/sites/default/files/rise_of_the_datavores.pdf, accessed 13 May 2015.

Barua, A., D. Mani, R. Mukherjee (2013), „Impacts of effective data on business innovation and growth", Chapter Two of a three-part study, University of Texas at Austin, www.businesswire.com/news/home/20100927005388/en/Sybase-University-Texas-Study-Reveals-Incremental-Improvement, accessed 20 May 2015.

Berners-Lee, T. (2007), "Q&A with Tim Berners-Lee", Bloomberg Business, 9 April, www.bloomberg.com/bw/stories/2007-04-09/q-and-a-with-tim-berners-leebusinessweek-business-news-stock-market-and-financial-advice, accessed 4 May 2015.

BLS-OES (2014), Occupational Employment Statistics, US Bureau of Labor Statistics, November.

Big Data Startups (2013), "Walmart is making big data part of its DNA", www.bigdata-startups.com/BigData-startup/walmart-making-big-data-part-dna/, last accessed 22 August 2014.

BITKOM and Fraunhofer (2014), "Industrie 4.0 – Volkswirtschaftliches Potenzial für Deutschland", www.bitkom.org/files/documents/Studie_Industrie_4.0.pdf, accessed 15 June 2015.

Bollier, D. (2010), *The Promise and Peril of Big Data*, The Aspen Institute, Washington, DC.

Bruner, J. (2013), "Defining the industrial Internet" O'Reilly Radar, 11 January, http://radar.oreilly.com/2013/01/defining-the-industrial-internet.html.

Brynjolfsson, E. and A. McAfee (2014), *The Second Machine Age – Work, Progress and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

Brynjolfsson, E., L.M. Hitt and H.H. Kim (2011), "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?", Social Science Research Network (SSRN), 22 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

Brynjolfsson, E. et al. (2008), "Scale without mass: Business process replication and industry dynamics", Harvard Business School Technology & Operations Mgt. Unit Research Paper No. 07-016, 20 September, http://dx.doi.org/10.2139/ssrn.980568.

CEBR (2012), "Data equity: Unlocking the value of big data", Centre for Economics and Business Research Ltd, London, www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf, accessed 15 April 2015.

CEPS (2014), "Shaping the integrated infrastructures of cities", presentation for the IEC CEPS workshop on "Orchestrating smart city efficiency", Centre for European Policy Studies, www.ceps.eu/system/files/article/2014/04/CEPS_IEC_Smart_City_3-Integrated_Infrastructures.pdf, accessed 15 May 2015.

Chick, S., S. Netessine and A. Huchzermeier (2014), "When big data meets manufacturing", *Knowledge*, INSEAD, 16 April, http://knowledge.insead.edu/operations-management/when-big-data-meets-manufacturing-3297, accessed 15 May 2015.

Christensen, C. M. (1997), *The Innovator's Dilemma*, Harvard Business School Press, Boston.

Corrado, C., C. Hulten and D. Sichel (2009), "Intangible capital and U.S. economic growth", Review of Income and Wealth, Series 55, No.3, September, www.conference-board.org/pdf_free/IntangibleCapital_USEconomy.pdf, accessed 15 May 2015.

Cowen, T. (2013), *Average is Over: Powering America Beyond the Age of the Great Stagnation*, Dutton Adult.

Croft, A. (2014), "Obama says U.S. needs to win back trust after NSA spying", *Reuters*, 25 March, www.reuters.com/article/2014/03/25/us-usa-security-obama-spying-idUSBREA2O18T20140325, accessed 15 May 2015.

Duhigg, C. (2012), "How companies learn your secrets", *New York Times*, 16 February, www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Dumbill, E. (2012), "Microsoft's Plan for Big Data", in *O'Reilly Planning for Big Data*, http://oreilly.com/data/radarreports/planning-for-big-data.csp, accessed 15 May 2015.

EC (2010), "Riding the wave: How Europe can gain from the rising tide of scientific data", Final report by the High-level Expert Group on Scientific Data, European Commission, October, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf, accessed 15 May 2015.

Economist Intelligence Unit (2014), "Networked manufacturing: The digital future", *Economist Intelligence Unit* sponsored by Siemens, 7 July, www.economistinsights.com/technology-innovation/analysis/networked-manufacturing.

Economist Intelligence Unit (2012), "The deciding factor: big data & decision making", *Economist Intelligence Unit* commissioned by Capgemini, 4 June, www.capgemini.com/insights-and-resources/by-publication/the-deciding-factor-big-data-decision-making/.

Elliott, S. (2014), "Anticipating a Luddite revival", *Issues in Science and Technology*, Spring, pp. 27-37, http://issues.org/30-3/stuart/, accessed 25 May 2015.

EOP (2014), "Big Data: Seizing opportunities, preserving values", Executive Office of the President, United States, www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, accessed 15 May 2015.

ESG (2012), "Boiling the ocean of control points in the Hadoop big data market", Enterprise Strategy Group, www.esg-global.com/blogs/boiling-the-ocean-of-control-points-in-the-hadoop-big-data-market/, accessed 24 May 2015.

Esmeijer, J., T. Bakker, and S. de Munck (2013), "Thriving and surviving in a data-driven society", TNO, 24 September, http://publications.tno.nl/publication/34610048/xcv74S/TNO-2013-R11427.pdf.

Ford, M. (2009), *The lights in the tunnel: Automation, accelerating technology and the economy of the future*, Acculant Publishing.

Frey, C.B. and M. Osborne (2013), *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, Oxford Martin School, University of Oxford.

Frischmann, B.M. (2012), *Infrastructure: The Social Value of Shared Resources*, Oxford University Press.

Frischmann, B.M., M.J. Madison, and K.J. Strandburg (2014), *Governing Knowledge Commons*, Oxford University Press.

Gerdon, S. and C. Reimsbach-Kounatze (2015), "Data-driven innovation for development: Unleashing 'big data' for inclusive growth", *OECD Digital Economy Papers*, OECD Publishing, Paris, forthcoming.

Guellec, D. and S. Wunsch-Vincent (2009), "Policy responses to the economic crisis: Investing in innovation for long-term growth", *OECD Digital Economy Papers*, No. 159, OECD Publishing, Paris, http://dx.doi.org/10.1787/222138024482.

Hardin, G. (1968), "The tragedy of the commons", Science (AAAS), Vol. 162, No. 3859, pp. 1243-1248, http://dx.doi.org/10.1126/science.162.3859.1243. PMID 5699198.

Harford, T. (2014), "Big data: Are we making a big mistake?", *Financial Times*, 28 March, www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html.

Hill, L. (2012), "How Target figured out a teen girl was pregnant before her father did", Forbes, 16 February, www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

IDC (2012), "Worldwide big data technology and services 2012-2015 forecast", IDC, March.

Jasperneite, J. (2012), "Was hinter Begriffen wie Industrie 4.0 steckt", computer-automation.de, 19 December, www.computer-automation.de/steuerungsebene/steuern-regeln/artikel/93559/0/.

Kramer, A.D.I., J.E. Guillory, and J.T. Hancock (2014), "Experimental evidence of massive-scale emotional contagion through social networks", Proceedings of the National Academy of Science of the United States of America (PNAS), Vol. 111, pp. 8788-8790, www.pnas.org/content/111/24/8788.full, accessed 2 June 2015.

Lane, J. et al. (eds.) (2014), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press.

Letouzé, E. and J. Jütting (2014), "Official statistics, big data and human development: Towards a new conceptual and operational approach", *Data-Pop Alliance White Papers Series*, 17 November, http://static1.squarespace.com/static/531a2b4be4b009ca7e474c05/t/546984d1e4b054b6f2656ac5/1416201425149/WhitePaperBigDataOffStatsNov17Draft.pdf.

Levy, F. and R.J. Murnane (2013), "Dancing with robots: Human skills for computerized work", Third Way, 1 June, http://dusp.mit.edu/uis/publication/dancing-robots-human-skills-computerized-work, accessed 2 June 2015.

Lodefalk, M. (2010), "Servicification of manufacturing - Evidence from Swedish firm and enterprise group level data", Working Papers No. 2010:3, Örebro University, School of Business, http://ideas.repec.org/p/hhs/oruesi/2010_003.html.

Mandel, M. (2013), "The data economy is much, much bigger than you (and the government) think", *The Atlantic*, 25 July, www.theatlantic.com/business/archive/2013/07/the-data-economy-is-much-much-bigger-than-you-and-the-government-think/278113/.

Mandel, M. (2012), "Beyond goods and services: The (unmeasured) rise of the data-driven economy", Progressive Policy Institute, 10 April, www.progressivepolicy.org/2012/10/beyond-goods-and-services-the-unmeasured-rise-of-the-data-driven-economy/.

Mayer-Schönberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray, London.

Merriam-Webster (2014), "Crowdsourcing", *Merriam-Webster.com,* www.merriam-webster.com/dictionary/crowdsourcing, accessed 24 September 2014.

McKinsey Global Institute [MGI] (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company, June, www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx, accessed 15 June 2013.

MIC (2013), "Information and Communications in Japan", White Paper, Ministry of Internal Affairs and Communications, Japan.

Moore, F. (1993), "Predators and prey: A new ecology of competition", *Harvard Business Review*, May-June, http://blogs.law.harvard.edu/jim/files/2010/04/Predators-and-Prey.pdf, accessed 15 June 2013.

Naughton, J. (2015), "Don't trust your phone, don't trust your laptop – this is the reality that Snowden has shown us", *The Guardian*, Opinion, 8 March, www.theguardian.com/commentisfree/2015/mar/08/edward-snowden-trust-phone-laptop-sim-cards.

Nordregio (2014), "Urban planning and big data – Taking LUTi models to the next level?", www.nordregio.se/en/Metameny/Nordregio-News/2014/Planning-Tools-for-Urban-Sustainability/Reflection/, accessed 19 September 2014.

Noyes, K. (2014), "Cropping up on every farm: Big data technology", Fortune, 30 May, http://fortune.com/2014/05/30/cropping-up-on-every-farm-big-data-technology/.

OECD (2015a), *Digital Economy Outlook 2015*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264232440-en.

OECD (2015b), *Inquiries into Intellectual Property's Economic Impact*, OECD Publishing, Paris, forthcoming.

OECD (2015c), *OECD Science, Technology and Industry Scoreboard 2015*, OECD Publishing, Paris, forthcoming.

OECD (2014a), 2014 Ministerial Council Statement, *Resilient Economies and Inclusive Societies – Empowering People for Jobs and Growth*, OECD Publishing, Paris, 07 May, www.oecd.org/mcm/2014-ministerial-council-statement.htm.

OECD (2014b), *Addressing the Tax Challenges of the Digital Economy*, OECD/G20 Base Erosion and Profit Shifting Project, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264218789-en.

OECD (2014c), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264221796-en.

OECD (2014d), "The economics of transition to Internet Protocol version 6 (IPv6)", *OECD Digital Economy Papers*, No. 244, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxt46d07bhc-en.

OECD (2013a), *Supporting Investment in Knowledge Capital, Investment and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-7-en.

OECD (2013b), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'big data'", *OECD Digital Economy Papers*, No. 222, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k47zw3fcp43-en.

OECD (2013c), "Strengthening health information infrastructure for health care quality governance: Good practices, new opportunities and data privacy protection challenges", OECD Health Policy Studies, OECD Publishing, Paris.

OECD (2013d), "Building blocks for smart networks", OECD Digital Economy Papers, No. 215, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k4dkhvnzv35-en.

OECD (2013e), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, OECD Publishing, Paris, 11 July, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2013f), *How's Life? 2013*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264201392-en.

OECD (2012), *OECD Internet Economy Outlook 2012*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264086463-en.

OECD (2011), *Society at a Glance 2011*, OECD Publishing, Paris, http://dx.doi.org/10.1787/soc_glance-2011-en.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, OECD Publishing, Paris, 30 April, C(2008)36, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2006), Recommendation of the Council concerning Principles and Guidelines for Access to Research Data from Public Funding, OECD Publishing, Paris, 14 December, www.oecd.org/sti/sci-tech/38500813.pdf.

OECD and Eurostat (2005), *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*, OECD Publishing, Paris.

OECD and FAO (2012), *OECD-FAO Agricultural Outlook 2012-2021*, OECD Publishing, Paris.

Orrick (2012), "The big data report", Orrick, www.cbinsights.com/big-data-report-orrick, accessed 19 May 2015.

Pariser, E. (2012), *The Filter Bubble: How the New Personalised Web is Changing What We Read and How We Think*, Penguin Books, April.

PCAST (2014), "Big data and privacy: A technological perspective", President's Council of Advisors on Science and Technology, United States, May, www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf, accessed 19 May 2015.

Piatetsky, G. (2014), "Did Target really predict a teen's pregnancy? The inside story", KDnuggets, 7 May, www.kdnuggets.com/2014/05/target-predict-teen-pregnancy-inside-story.html.

Pingdom (2013), "The top 100 web hosting countries", 14 March, available at http://royal.pingdom.com/2013/03/14/web-hosting-countries-2013/, accessed 19 May 2015.

Pole, A. (2010), "How Target gets the most out of its guest data to improve marketing ROI", Presentation at the 2010 PAW - the Predictive Analytics World Conference, 19-20 October, www.rmportal.performedia.com/node/1373, accessed 19 May 2015.

Reimsbach-Kounatze, C. (2015), "The proliferation of data and implications for official statistics and statistical agencies: A preliminary analysis", *OECD Digital Economy Papers*, OECD Publishing, Paris, http://dx.doi.org/10.1787/5js7t9wqzvg8-en.

Rifkin, J. (2014), *The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism*, Palgrave Macmillan.

Rose, C. (1986). "The comedy of the commons: Custom, commerce, and inherently public property", The University of Chicago Law Review, pp. 711-81, http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=2827&context=fss_papers, accessed 19 May 2015.

ScienceDaily (2013), "Big Data, for better or worse: 90% of world's data generated over last two years", 22 May, www.sciencedaily.com/releases/2013/05/130522085217.htm, accessed 11 April 2015.

Squicciarini, M. and M. Le Mouel (2012), "Defining and measuring investment in organisational capital: Using US microdata to develop a task-based approach", OECD Science, Technology and Industry Working Papers, No. 2012/5, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k92n2t3045b-en.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of government in a digital age", Computer and Communications Industry Association, October, www.ccianet.org/CCIA/files/ccLibraryFiles/Filename/000000000086/govtcomp_report.pdf, accessed 10 October 2013.

Tambe, P. (2014), "Big data investment, skills, and firm value", Management Science, forthcoming, http://ssrn.com/abstract=2294077, accessed 10 June 2015.

The Economist (2012), "High-frequency trading: The fast and the furious", The Economist, 25 February, www.economist.com/node/21547988.

The Economist (2010a), "Data, data everywhere", 25 February, www.economist.com/node/15557443, accessed 10 June 2015.

UN Global Pulse (2014), "Mining Indonesian tweets to understand food price crises", United Nations Global Pulse, February, www.unglobalpulse.org/sites/default/files/Global-Pulse-Mining-Indonesian-Tweets-Food-Price-Crises%20copy.pdf, accessed 19 May 2015.

UN Global Pulse (2012), "Big data for development: Opportunities & challenges", United Nations Global Pulse, May, www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf, accessed 10 June 2015.

Vennewald, L. (2013), "A Quantum Leap Forward for Logistics Players", T-Systems Best Practices, October, www.t-systems.pt/umn/short-messages-about-customer-projects-innovations-and-solutions-for-cloud-computing-and-big-data-t-systems/1162170_1/blobBinary/Best-Practice_03-2013_News_EN.pdf, accessed 19 May 2015.

Wall, M. (2014), "Big data: Are you ready for blast-off?", BBC, 4 March, www.bbc.com/news/business-26383058, accessed 19 May 2015.

WEF (2012), "Big data, big impact: New possibilities for international development", World Economic Forum, www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf, accessed 6 June 2013.

Chapter 2

Mapping the global data ecosystem and its points of control

In exploring the rapidly evolving data ecosystem, this chapter enumerates the key actors, their main technologies and services, and their business and revenue models. It uses a layer model to identify these actors as well as strategic points of control in the system. It goes on to discuss the interaction among actors, analysing in particular the relation between competition and collaboration for DDI, and how this "co-opetition" translates in terms of horizontal and vertical dynamics. The chapter analyses the degree to which data ecosystems are open, global and interconnected. Finally, it looks at the implications of DDI for global value chains (GVCs) and trade, taxation, and competition.

The great thing about big data is that there's still plenty of room for new blood, especially for companies that want to leave infrastructure in the rearview mirror. (Harris, 2012)

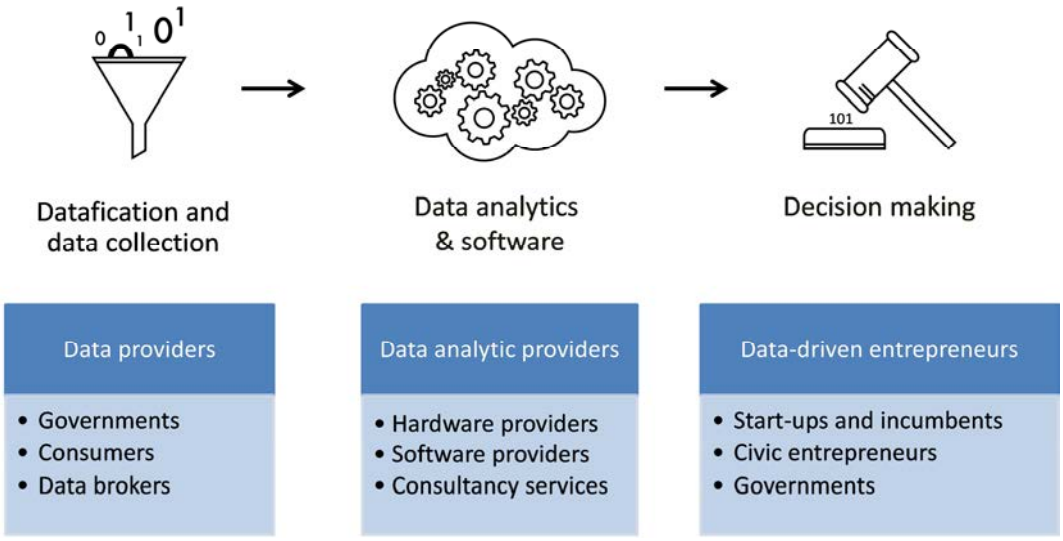
I look at this audience, and I look at VMware and the brand reputation we have in the enterprise, and I find it really hard to believe that we cannot collectively beat a company that sells books. (VMware's President and COO Carl Eschenbach, VMware Partner Exchange conference, February 2013)

Data-driven innovation – DDI, introduced in Chapter 1 of this volume – refers to the use of data and analytics to improve or foster new products, processes, organisational methods and markets. It is the concrete fulfilment of the value creation process along the data value cycle (see Figure 1.7 in Chapter 1), embarked upon in order to reach a specific goal, tackle a problem, or grasp an opportunity for which data analytics could provide (a part of) the solution. Each specific goal will require an organisation (or a consortium of organisations) to organise a value creation process along the data value cycle. It is likely that for many of the steps in this process, organisations will have to involve third parties around the world, because they lack experience, technological resources and/or talent to deal with the multidisciplinary aspects of data and analytics on their own. The resulting global value chain (GVC) is in most cases specifically tailored towards the goal that is being pursued. The combined effect is that a global data ecosystem is emerging in which, more than ever before, data and analytic services are traded and used across sectors and across national borders. For the information and communication technology (ICT) industry this represents a USD 17 billion business opportunity for 2015, with an estimated market growth of more than 40% on average every year since 2010 (see IDC, 2012; Kelly, 2013).[1]

Better analysis of both the economic and societal impacts of DDI requires a deeper understanding of the complexity and dynamics of the emerging global data ecosystem – including the interaction between the actors, their technologies and their business models, and the dynamics that structure this ecosystem. The concept of an ecological approach to describe business environments chosen for the analysis in this chapter was introduced by Moore (1993) to describe how companies should not be viewed as members of a single industry "[…] but as part of a business ecosystem that crosses a variety of industries." In these ecosystems, collaborative arrangements of firms combine their individual offerings to create coherent, customer-facing solutions (Adner, 2006). This is an appropriate perspective with which to explore the dynamics of networks of human and non-human actors, that have started to form around specific outcomes of DDI, and that may gradually link together into an all-encompassing global data ecosystem.

This chapter analyses that ecosystem, using the data value cycle introduced in Chapter 1 as a framework for identifying the different types of companies and services competing within it (Figure 2.1). The chapter also analyses other factors affecting the functioning of the data ecosystem such as key technologies, business models, and coalitions/alliances that are forming. By mapping actors and their technologies and business models using a "follow the data" approach along the data value cycle, the chapter reveals links across sectors, potential points of control in the data ecosystem, and their gatekeepers that mediate the interactions that shape the ecosystem. Newly forming GVCs will point to new actors and emerging horizontal data markets. This chapter builds on a rich mix of comprehensive expert interviews, case studies and workshops, all conducted by TNO, Netherlands Organisation for Applied Scientific Research.[2]

Figure 2.1. **Main phases of the data value cycle with their key types of actors**

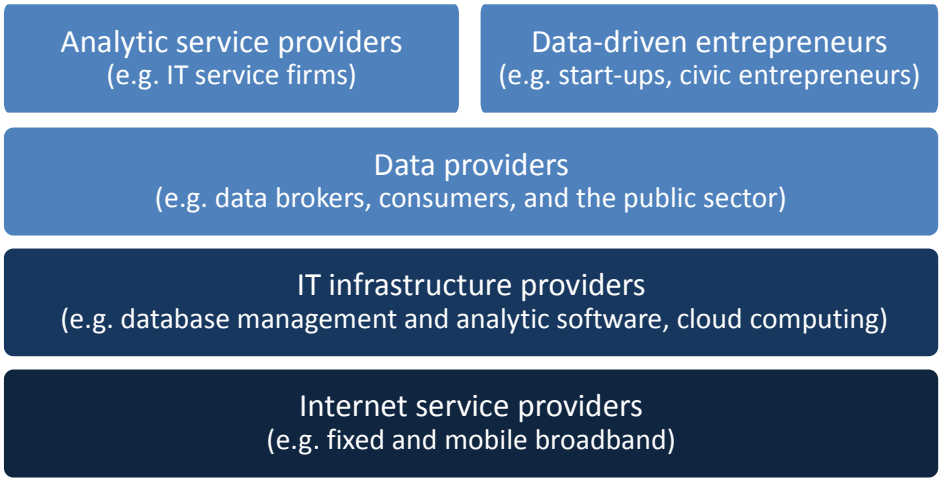


2.1. The key actors and their main technologies, services and business models

With the increase in data generated, collected and stored, and a greater variety of information that can be extracted from this data, companies in the data ecosystem are mushrooming. The number of actors, goods and services, and technologies and business models that shape the data-driven constellations creates a rapidly evolving data ecosystem. Describing this data ecosystem is problematic due to its constant evolution. It is a dynamic field as new technologies and practices are constantly being developed, driven largely by traditional information technology (IT) businesses in infrastructure and business analytics (e.g. IBM, Oracle, SAP and Microsoft) and a vast array of new start-ups. The increasingly active role of non-traditional data companies in the data ecosystem is also remarkable. One of the most telling examples is Amazon. Its role as a big data powerhouse is clearly illustrated by this chapter's opening quotation from VMware's President and COO Carl Eschenbach in response to Amazon's rise at the expense of VMware (Assay, 2013).

There are various depictions of the data ecosystem that position the different types of actors. Turck (2014), for example, illustrates the different clusters of businesses based on a detailed typology of services, products and technologies, including: i) the underlying core technologies, such as Hadoop and Cassandra; ii) the IT infrastructure (e.g. storage and computing); iii) the analytical tools (e.g. "R"); and iv) domain-specific applications. In this chapter, the data ecosystem is seen as a combination of layers of key roles of actors, where the underlying layers provide goods and services to the upper layers (Figure 2.**2**).

Figure 2.2. **The data ecosystem as layers of key roles of actors**



The following sections describe the different roles of actors, their technologies and services, and their main business models (including their revenue models) in more detail. What is provided is a generalised overview of the field as of 2014, with the understanding that as new actors enter and technologies evolve, so will the relative positions of the various players. The different roles include:

1. *Internet service providers* – Internet service providers form the backbone of the data ecosystem through which data are exchanged.
2. *IT infrastructure providers* – The second layer includes IT (hardware and software) infrastructure providers that are offering data management and analysis tools and critical computing resources, including but not limited to data storage servers, database management and analytic software, and most importantly cloud computing resources.
3. *Data (service) providers* – The third layer includes i) data brokers and data marketplaces that are selling their data across the economy, ii) the public sector with its open data initiatives (see Chapter 10 of this volume), and – last but not least – iii) consumers that are *actively* contributing their data to the data ecosystem increasingly as well, thanks to new services provided by innovative businesses but also through data portability initiatives (Chapters 4 and 5).
4. *Data analytic service providers* – The fourth layer includes businesses that provide data aggregation and analytic services, mainly to business customers. This also includes data visualisation services.
5. *Data-driven entrepreneurs*[3] – These entrepreneurs build their innovative businesses based on data and analytics available in the ecosystem. Their efforts result in DDI for science and research (see Chapter 7), health care (Chapter 8), smart cities (Chapter 9), and public service delivery (Chapter 10).

Before looking at each type of actor separately and in more detail, it is important to acknowledge five important characteristics of the data ecosystem that can only partially be reflected in an analysis based solely on the typology of the key actors.

1. Figure 2.2 implicitly suggests that firms can only be assigned to one particular role. However, a closer look at the business model of the key actors reveals that many businesses will typically play multiple roles. Internet service providers (ISPs), for example, are increasingly using data analytic services to manage their networks (OECD, 2014a),[4] but also to generate data on, for example, communication patterns that are offered to third parties. In the latter case, ISPs are acting as data service providers. For example, the French mobile ISP Orange uses its Floating Mobile Data (FMD) technology to collect mobile telephone traffic data that are anonymised and sold to third parties, including government agencies and traffic information service providers. Furthermore, IT infrastructure providers may furnish the full stack of hardware and software solutions needed for data analysis including through cloud-based services, on top of which access to third parties' data are also provided. Microsoft's cloud service Microsoft Azure, for example, is provided with the Azure Data Marketplace, where users can access data sets provided by third parties. These multiple roles of actors not only challenge measurement efforts for statistical purposes (see Box 2.1), but also point to the significant share of vertically integrated companies such as Google and Microsoft, and most importantly to the dual nature of many actors in the data ecosystem as users and producers of data and analytics.[5] This dual role suggests that the data ecosystem is a logical continuation of the Web 2.0.[6]
2. Figure 2.2 does not reflect the inherently global nature of the data ecosystem. The data ecosystem involves cross-border data flows due to the activities of key global actors and the global distribution of technologies and resources used for value creation. In particular, ICT infrastructures used to perform data analytics, including the data centres and software, will rarely be restricted to a single country, but will be distributed around the globe to take advantage of several factors; these can include local work load, the environment (e.g. temperature and sun light), and skills and labour supply (and costs). Moreover, many data-driven services developed by entrepreneurs "stand on the shoulders of giants" who have made their innovative services (including their data) available via application programming interfaces (APIs), many of which are located in foreign countries (see Chapter 3).
3. Related to the global nature of the data ecosystem is the missing representation in Figure 2.2 of the "cytoplasm" that lies between the layers of the data ecosystem and that enables the smooth *interoperability* of the different types of actors, their technologies, and services. Open Internet standards such as TCP/IP and HTTP have been and still are crucial for the global data ecosystem, which relies heavily on the open Internet for its functioning. In addition, the reuse of data and of data-driven services underlines the importance of (open) standards related to e.g. APIs and data formats (including meta-information and data), which are a requirement for the interoperability of data-driven services and the portability of data across these services.

Box 2.1. **The challenge in measuring "big data"-related industries**

In 2012, the OECD[1] undertook efforts to measure the value-added of big data-related activities identified from a National Accounts (NA) perspective. The work, based on a questionnaire submitted to 25 countries,[2] highlighted issues in attempting to derive estimates of the size of these activities.

Big data-related industries were identified as those industries collecting, processing and diffusing digital data. These industries were then retraced in the international classification of economic activities (ISIC) at the finest available level of disaggregation (i.e. at the class level, corresponding to four digits). This way of proceeding aims at producing estimates that would eventually be reliable and comparable with other NA aggregates and across countries. The result was an operational definition of big data-related industries. With reference to the latest release of ISIC (Rev.4), these activities would fall into the classes 5812: *Publishing of directories and mailing lists*; 5819: *Other publishing activities*; 6311: *Data processing, hosting and related activities*; and 6312: *Web portals*. ISIC Rev.4 offers a finer classification of information activities with respect to the previous Rev.3.1, and groups them together under Section "J". Nonetheless, the correspondence is far from perfect. Indeed, the above ISIC classes also include activities outside the data industry aggregate, such as web hosting (under class 6311), or publishing (under 5812 and 5819).

Overall, the *share* of value-added of digital data industries in the United States alone appeared to be much higher than in all the other countries considered together. This result was partly attributable to the likely higher development of these industries in the United States, but also suggested the need for a more comprehensive assessment when undertaken by NA. Additionally, data for some countries did not include those of all industries in the *digital data* aggregate. In particular, a number of respondents could not provide information on activities in the *publishing* industries classes. Regarding the relative weight of big data-related activities for other countries, one could observe that some "plausible" figures also deviated from what would be expected. Employment figures showed a pattern similar to those observed for value-added, with several data missing and wide country variations.

This first exploration provided a very preliminary perspective on the size of big data-related activities. Given the tiny size of the aggregate and the high variability across countries, these results were far from reliable, and not simply because of lack of coverage. In general, for all countries but the United States the source of data lay within the domain of Structural Business Statistics. This implied that in principle, data would include only businesses whose *main activity* fell in a given class – excluding those performing, say, data processing as a secondary activity (multi-product enterprises) or those in which data are instrumental to their main activity ("own account", as for e.g. financial firms). The coverage of publishing activities (ISIC classes 5812 and 5819) was not complete for all responding countries, while in some cases only aggregates for data processing and hosting (6311) and web-related activities (6312) were available. The above elements led to substantial underestimation of the size of big data-related activities in some countries. However, other elements of as yet unknown magnitude also influenced measurement in the opposite direction, such as the weight of activities included in the above classes that were not related to big data.

Box 2.1. **The challenge in measuring "big data"-related industries (cont.)**

At this stage, results based on NA are therefore considered an indication of the work that remains to be done, rather than a first approximation of the size of the industry. Relevant improvements could be achieved by harmonising the information produced, with specific attention paid to big data-related activities in statistical production. A more precise definition of industry boundaries, such as at the six-digit NAICS (the North American Classification), could be envisaged in the future. In principle, where estimates are to be performed by NA, this could also include secondary and own-account activities; recent evidence (including the study by Bakhshi and Mateos–Garcia, 2012) shows these to be a relevant component of the digital data aggregate. Practically speaking, NA estimates based on integration techniques are not immediately feasible. A necessary prerequisite – and at the same time a good result in itself – would be to achieve more precision in business statistics. In many cases the infrastructure is available, as published data are already based on five- or six-digit information; it could be reinforced at Kind of Activity Unit / Establishment level, be linked to product classifications, and be extended to working hours. In this respect, the exclusion of some activities such as webhosting and certain printing activities is highly recommended. Also, a thorough check (and agreement) on which activities are (to be) reported in individual classes would be beneficial, with the possible exclusion of specific activities which "by definition" are in the big data domain but seem to be substantively different. These operations, together with closer monitoring of data robustness by national statistic offices (NSOs), could help improve data quality. However, they may require significant time and efforts by NSOs.

1. The work was undertaken by the OECD Working Party on Measurement and Analysis of the Digital Economy (WPMADE, formerly the Working Party on Indicators for the Information Society, WPIIS) and presented at the December 2012 meeting of the working party.

2. Twenty-on of the responses where provided with detailed data set in ISIC Rev.3.1 and ISIC Rev.4.

4. Figure 2.2 does not reflect the fact that the data ecosystem relies on a variety of business models that may not necessarily be linked to the role of the actors within the ecosystem, but rather to the market segment targeted by these actors. To recall, a business model specifies the value proposition of a business, including its key activities and the goods and services (i.e. products) it offers. The business model also specifies the targeted market segment and most importantly the *revenue models* that describe how the business turns the value of their products into revenues. Analysis of business models of companies in the data ecosystem suggests that the selection of revenue models in the ecosystem mainly depends on whether the business model focuses on business to business (B2B) or business to consumer (B2C) offers.[7] In addition, businesses in the data ecosystem use a diversity of revenue models, some of which are often combined to maximise revenues (see Box 2.2). The resulting complexity of the mechanisms through which revenues are generated has led to a number of policy challenges, such as the challenge of value attribution (with implications for taxation) and the challenge faced by competition authorities in defining the relevant markets. Both policy issues are discussed further below.

Box 2.2. **The diversity of revenue models in the data ecosystem**

Businesses in the data ecosystem use a diversity of revenue models, some of which are often combined to maximise revenues. The most common models include the following.

Freemium – The term "freemium" is a portmanteau of the words "free" and "premium". The *freemium* revenue model, one of the most dominant in the data ecosystem, seems to be particularly attractive to start-ups: products are provided free of charge, but money is charged for additional, often proprietary features of the product (i.e. *premium*). The freemium revenue model is often combined with the advertising-based revenue model for B2C offers, where the free product is offered with advertisement while the premium offer is advertisement-free.

Advertisement – Advertisement is most frequently used for B2C offers: products are offered free of charge or with a discount to users in exchange for required viewing of paid-for advertisements (OECD, 2014d) Increasingly, advertisement is provided based on the profile and/or location of the consumers. Advertisement-based revenue models are also used in multi-sided markets together with *cross subsidies*, where a service is provided for free or at a low price on one side of the market, but subsidised with revenues from other sides of the market.

Subscription – Subscription-based revenue models are by far the models most frequently used in the data ecosystem, for B2B offers in particular (among all the B2B business models of start-ups analysed by Hartmann et al. (2014), for example, 98% were subscription based). Examples of subscription-based models include regular (daily, monthly or annual) payments for access to the Internet, as well as access to digital content including data, news, music, video streaming, etc. The category also includes regular payments for software services and maintenance, hosting and storage, and customer "help" services. Subscription-based revenue models are often combined with the *freemium revenue model*, where the premium product is provided with a subscription (see above).

Usage fees – Usage fees are the second most frequently used revenue model used by start-ups in the data ecosystem. They are also a prominent revenue model for B2B offers. Usage fees are typically charged to customers for use of a particular (online) service – including most offers that are provided "as-a-Service" (XaaS), such as cloud computing based services for example (see section on IT infrastructure providers). These services are offered through a *pay-as-you go model*, where usage fees are charged for the actual use of the service.

Selling of goods (including digital content) – Asset sale is still used in the data ecosystem, mainly by IT infrastructure providers. But it is also used by service platform providers that sell sensor-equipped smart devices (including smartphones, smart meters and smart cars) as a source for generating data and delivering value-added services. Furthermore, it includes *pay-per-download* revenue models where users pay per item of download. These could include, for instance, data sets or other digital content such as e-books, videos, apps, games and music.

Selling of services – This revenue model includes the provision of traditional B2B services such as IT consultancy services, software development and maintenance and helpdesk support. It also includes a wide range of long-term B2B services provided by Internet intermediaries such as web hosting, domain registration, and payment processing. It thus overlaps with the revenue models that are based on subscriptions and usage fees often used for IT service contracts.

Licensing – This revenue model is often used to generate revenues from intangible assets that are protected through intellectual property rights (IPRs), such as patents and copyrights. Licensing may thus be used to monetise software and software components including algorithms, libraries and APIs. It may also be used for databases. However, evidence suggests that licensing may not be an essential revenue model for start-ups, although it may be an important for well-established IT providers including in particular software companies (among the 100 start-ups analysed by Hartmann et al. (2014), none has indicated licensing as a source of revenue).

Box 2.2. **The diversity of revenue models in the data ecosystem (cont.)**

Commission fees – This is mainly used in B2C markets by intermediaries that use data analytics to better match supply and demand. Payment often will be calculated on the basis of a percentage of the price of products supplied, and it will only be obtained when successfully matching supply and demand – that is, when successfully providing businesses with customers.

5. Finally, although most illustrations of the data ecosystem such as Figure 2.2 provide an extensive and useful overview of the most relevant roles of actors in the ecosystem, they tend to be strongly ICT sector biased. They describe data-related technologies, the various types of data-related products and services, and the companies that provide them. However, the analyses conducted by TNO (2013) and in other chapters of this volume strongly suggest that DDI is not just a technological (ICT supply-side) challenge. DDI also presents serious demand-side challenges: working processes, attitudes, changes in management and human resource (HR) policy. However, the services or products that support these organisational challenges are rarely represented, and often also too complex to be fully represented, in a simplified model of the data ecosystem such as Figure 2.2. Legal consultation for example is very important, especially for organisations that deal with personal data, and this kind of service is often provided by external legal advisors.

It should therefore be acknowledged that in focusing solely on technological aspects, the analysis of the data ecosystem presented in this chapter only accounts for a relatively small share of all the interactions and relationships within the data ecosystem. Any assessment of the ecosystem – in particular, quantitative assessment of its total market size – that does not consider this limitation risks underestimating its full size and impact.

Internet service providers

In general, Internet service providers (ISPs) build and operate networks, typically at the regional level. They grant subscribers (businesses and consumers) access to the Internet through physical transport infrastructure as they have the equipment and telecommunication network required for a point-of-presence on the Internet. This is necessary to allow users to access content and services on the Internet and content providers to publish or distribute data and information online (OECD, 2011a). ISPs thus help build the foundation of the data ecosystem as they provide local, regional and/or national (fixed and mobile) broadband coverage, or deliver backbone services for other ISPs.

Some ISPs are extending their product offer with for example web hosting, web-page design and consulting services related to networking software and hardware (OECD, 2011a). In this case, well-established ISPs can benefit from their established reputation to place themselves in new markets such as the IT service market (including e.g. cloud computing) in which consumers' trust plays an important role (Koehler, Anandasivam and Dan, 2010). Since 2010, Telefonica, Orange and Deutsche Telekom have launched cloud computing services targeting in particular small and medium-sized enterprises (SMEs) (Arthur D. Little, 2013). Some ISPs are going further up the value chain by providing data and analytic services. For example, the French mobile ISP Orange is acting as a data service provider by using its *Floating Mobile Data* (FMD) technology to collect mobile telephone traffic data; these determine speeds and traffic density at a given point in the road network, and deduce travel time or the formation of traffic jams. The

anonymised mobile telephone traffic data are sold to third parties, including government agencies, to identify "hot spots" for public interventions, but also to private companies such as Mediamobile, a leading provider of traffic information services in Europe.[8]

Another example is Telefónica, which in 2012 launched its new "big data business unit", Telefónica Dynamic Insights. This business unit, based in the United Kingdom, operates as an analytic service provider with the goal of providing companies and governments around the world with analytical insights based on mobile network and machine-to-machine (M2M) data. Its first product, *Smart Steps*, uses "anonymised and aggregated mobile network data to enable companies and public sector organisations to measure, compare, and understand what factors influence the number of people visiting a location at any time" (Telefónica, 2012).

The subscription model is the prevalent revenue model in the majority of OECD countries in which ISPs act as traditional Internet service providers. ISPs mostly charge a periodic – daily, monthly or annual – fee to subscribe to an unlimited service (OECD, 2011a). Other revenue models include those prepaid-based, or a combination of both subscription and prepaid. Prepaid models are commonly used by ISPs that meter their services, for example when mobile Internet access is offered. The price paid by the consumer is based on actual usage rates or a monthly subscription fee, with an additional amount charged for a data package (OECD, 2011a).

However, ISPs are currently debating whether the flat rate model will still be applicable in the future. Some ISPs have proposed differentiating among classes of Internet traffic (e.g. gold, silver bronze) or dedicating specific broadband capacity to certain applications. These plans are motivated by the rise of Internet traffic volume in particular due to the increased usage of video. Helping drive that increase are online streaming, such as Netflix offers in the United States and other countries, and online television, such as the BBC iPlayer in the United Kingdom and the Swedish company Magine TV that offers its service in Sweden, Germany and Spain (van der Berg, 2014). It has been argued that, if investments in networks continue to be made, the growth in traffic will not overwhelm networks since the growth rate of data traffic is strong but decreasing in relative terms (OECD, 2014a). In addition, ISPs appear to have been mostly unsuccessful in promoting a discriminatory pricing scheme. One reason put forward by content providers for not purchasing these services is, that their impact is mostly unknown as the ISPs control only part of the network. Furthermore, in a competitive market content providers may judge that ISPs will upgrade their networks when quality degrades to remain competitive with other ISPs (van der Berg, 2014).

IT infrastructure providers

The market for IT infrastructure comprises providers of both hardware and software. But most important for DDI are providers of databases and related technologies and services (management, security, transport, storage). These include in particular providers of platforms for distributed parallel data processing – such as Hadoop, which has almost become the standard technology to deal with more complex, unstructured large-volume data sets (Box 2.3). The importance of databases and related technologies and services is also reflected in estimates by IDC (2012), which suggest that "big data technology and services" will grow from USD 3 billion in 2010 to USD 17 billion in 2015. This represents a compound annual growth rate (CAGR) of almost 40%. Data storage technologies and services are estimated to be the fastest growing segment, followed by networking, and IT services, which explains the increasing role of IT equipment firms in this relatively new market (see section below on mergers and acquisitions, M&A).

Box 2.3. **Internet spillovers enabling data-driven innovation across the economy: The case of Hadoop**

Internet firms, in particular providers of web search engines, have been at the forefront in the development and use of techniques and technologies for processing and analysing large volumes of data. Google, in particular, inspired the development of a series of technologies after it presented *MapReduce*, a programming framework for processing large data sets in a distributed fashion, and *BigTable*, a distributed storage system for structured data, in a paper by Dean and Ghemawat (2004) and Chang et al. (2006) respectively. In 2006, the open source implementation of *MapReduce,* called *Hadoop,* emerged. Initially funded by Yahoo, *Hadoop* is now provided as an open source solution (under the Apache License) and has become the engine behind many of today's big data processing platforms. Beside Yahoo, Hadoop is ushering in many data-driven goods and services offered by Internet firms such as Amazon, eBay, Facebook, and LinkedIn. As mentioned above, even traditional providers of databases and enterprise servers such as IBM,[1] Oracle,[2] Microsoft[3] and SAP[4] have started integrating Hadoop and other related open source tools into their product lines, making them available to a wider number of enterprises including Walmart (retail), Chrevon (energy), and Morgan Stanley (financial services).

The key innovation of MapReduce is its ability "to take a query over a data set, divide it, and run it in parallel over many nodes" (Dumbill, 2010), often using (low-cost) commodity servers that can be distributed across different locations. This distribution solves the issue of data being too large to fit onto and to be processed by a single server. The data used for MapReduce also do not need to be relational or even to fit a schema, as is the case with the conventional (relational) SQL databases. Instead, unstructured data can be stored and processed. The standard storage mechanism used by Hadoop is therefore a distributed file system, called HDFS (Hadoop Distributed File System). On top of being distributed, HDFS is a fault tolerant file system that can scale up to dozens of petabytes (millions of gigabytes) of storage and can run with high data throughput on all major operating systems (Dumbill, 2010). However, other file systems are also supported by Hadoop, such as the Amazon S3 file system (used on Amazon's cloud storage service).

To simplify the use of Hadoop (and HDFS), additional open source applications have been developed or existing ones have been extended, some through the initiative of top Internet firms. HBase, for example, is an open source, non-relational (i.e. NoSQL) distributed database, also under the Apache Licence. HBase was modelled after Google's BigTable, and can run on top of HDFS or Hadoop. HBase is now, for example, currently used by Facebook for its Messaging Platform, which in 2010 had to support 15 billion person-to-person messages and 120 billion chat messages per month (Muthukkaruppan, 2010). Another example is Hive, an open source data warehouse infrastructure running on top of Hadoop, which was initially developed by Facebook to simplify management of structured data using a SQL-based language (HiveQL) for queries. Finally, analytical tools such as R, an open-source environment for statistical analysis, are increasingly being used in connection with Hive or Hadoop to perform big data analytics. The evidence suggests that R is becoming a more preeminent tool for data analytics (Muenchen, 2014).

The resulting ecosystem of big data processing tools can be described as a stylised stack of storage, MapReduce, query, and analytics application layers. Increasingly, the whole stack is provided as a cloud-based solution by providers such as Amazon (2009) and Microsoft (2011). One could argue along with Dumbill (2010) that this evolving stack has enabled and democratised big data analytics in the same way "the commodity LAMP stack of Linux, Apache, MySQL and PHP changed the landscape of web applications [and] was a critical enabler for Web 2.0" (Dumbill, 2010).

1. IBM is offering its Hadoop solution through InfoSphere BigInsights. BigInsights augments Hadoop with a variety of features, including textual analysis tools that help identify entities such as people, addresses and telephone numbers (Dumbill, 2012b).

Box 2.3. **Internet spillovers enabling data-driven innovation across the economy: The case of Hadoop (cont.)**

2. Oracle provides its Big Data Appliance as a combination of open source and proprietary solutions for enterprises' big data requirements. The appliance includes, among others, the Oracle Big Data Connectors to allow customers to use Oracle's data warehouse and analytics technologies together with Hadoop, the Oracle R Connector to allow the use of Hadoop with R, an open-source environment for statistical analysis, and the Oracle NoSQL Database, which is based on Oracle Berkeley DB, a high-performance embedded database.

3. In 2011, Microsoft began integrating Hadoop with Windows Azure, Microsoft's cloud computing platform, and one year later with Microsoft Server. It is providing Hadoop Connectors to integrate Hadoop with Microsoft's SQL Server and Parallel Data Warehouse (Microsoft, 2011).

4. In 2012, SAP announced its roadmap to integrate Hadoop with its real-time data platform SAP HANA and SAP Sybase IQ.

Until a few years ago, the nascent Hadoop space was dominated by a few products and their providers, such as the open-source Apache Hadoop distribution, the independent Hadoop distribution provider Cloudera and Amazon's Elastic Map Reduce (Harris, 2011a). But this space has rapidly become densely populated. According to Harris (2011b), the infrastructure market is already near its point of saturation:

> *The market for horizontally focused products is filling up fast with both start-ups and large vendors […] Yes, there's still room for start-ups to get in here, but the door looks to be closing fast. It's not just Hadoop, either; other techniques, from traditional data warehouses to, arguably, predictive analytics, all are nearing the saturation point in terms of vendors selling the core technologies (Hariss, 2011b).*

On the one hand, new independent Hadoop distribution actors emerged, such as Hadapt, HortonWorks (a Yahoo spin-off) and MapR. On the other hand, traditional infrastructure vendors that offer servers, storage and database technologies, moved into this space as well. IBM, EMC, Cisco, Oracle, HP and VMware have all adopted Hadoop in order to provide big data solutions to their customers – sometimes in partnership with the independent Hadoop distribution providers. They align their Hadoop products with the rest of their database and analytical offerings for business intelligence (Dumbill, 2012a).

Although Hadoop has proved to be very popular – especially for big, unstructured data challenges – classical (relational) database technologies are still important, as are next-generation massive parallel processing database technologies and their related analytical tools. Companies that provide these analytical platforms that combine databases and analytical tools are Vertica (owned by HP), Asterdata (owned by Teradata), SAP (with Hana), ParAccel, Attivo and Datastax, to name but a few. However, these products are often used in combination with Hadoop.[9]

In addition, data analytic solutions that help extract insights from data are also provided by IT providers in particular specialised analytic software companies such as SAS, The MathWorks, and RapidMiner. As highlighted in Chapter 3 of this volume, open source software (OSS) based on free software licences – such as the MIT License,[10] the BSD License,[11] the Apache License[12] and the GNU general public license (GPL v2 or v3)[13] are attracting an increasing number of business and consumer users. A well-known example is R mentioned in Box 2.3. A GPL licenced open-source environment for statistical analysis, R is increasingly used (sometimes together with Hadoop) as an

alternative to commercial packages such as SPSS (IBM) and SAS (Muenchen, 2014). The high popularity of R has even pushed traditional providers of commercial databases and enterprises servers (and competitors) such as IBM, Oracle, Microsoft, and SAP to integrate R (together with Hadoop) into their product lines, and to compete with the specialised analytic software companies.

Two trends in the business models of IT infrastructure providers can be observed. First, the business model of IT infrastructure providers is increasingly characterised by the *freemium* revenue model described in Box 2.2, where products are provided free of charge, but money is charged for additional, often proprietary, features (*premium*). This model commonly used in the open source software industry, play a major role in the data ecosystem – most likely because of the prominence of open source software solutions such as Hadoop and R, as described above. IT infrastructure vendors such as Hortonworks, Cloudera and MapR, for example, are providing at least one basic version their products for free. Revenues are then generated either based on premium versions of the product or based on value added complementary services. Hortonworks, for example, provides just one version of its Hadoop solution, called "Hortonworks Data Platform", at no cost to download. Around two-thirds of its revenues are generated based on annual subscription services contracts, which are the equivalent of maintenance and supports contracts customarily provided by virtually all businesses in the software industry (Kelly, 2013). The remaining third of the revenues is generated based on professional training services. Cloudera, as another example, provides its Hadoop solution "Cloudera's Distribution Including Apache Hadoop" (CDH) with a proprietary software component for free. The full version of the package "Cloudera Enterprise" is available for an annual for-pay subscription, however.

The second key trend that has substantially changed the business models of IT infrastructure providers is *cloud computing* (see Chapter 3 of this volume). Cloud computing has been described as "a service model for computing services based on a set of computing resources that can be accessed in a flexible, elastic, on-demand way with low management effort" (OECD, 2014c). Cloud computing can be classified into three different service models according to the resources it provides: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS) (see Chapter 3 for further information).[14]

As cloud-based services increasingly become viable alternatives to (parts of the) infrastructure, actors such as Amazon, Google and Microsoft, and other new entrants specialised in cloud computing, are continuously challenging the predominant business models of IT infrastructure providers. Cloud computing providers are also extending their services in the market; some offer not only data storage and management solutions as IaaS, but also data analytic solutions as either PaaS or SaaS. The key business model innovation of cloud computing providers is the fact that their services are offered through a pay-as-you go model (a usage fee-based revenue model), which enables cloud users to act more responsively to their needs and their customers' demand without much initial investment in IT infrastructure. That innovation lowers the entry barriers for start-ups and small and medium-sized enterprises (SMEs), but also for governments that cannot or do not want to make heavy upfront investments in ICTs; it consequently makes the markets more competitive and more innovative (see Chapter 3).

Data (service) providers

The stacks of technologies, analytics platforms, and applications are all tailored to process data, transforming them into valuable information and insights or otherwise actionable output. But at the heart of the current data ecosystem lies data, which some have characterised as the "life blood" or the "oil" of the ecosystem (see Chapter 4). The sections that follow discuss various groups of data providers. Their business model could be described in analogy to the cloud computing value proposition as Data-as-a-Service (DaaS; see Chen et al., 2012). Their revenue models however can vary significantly.

Data brokers

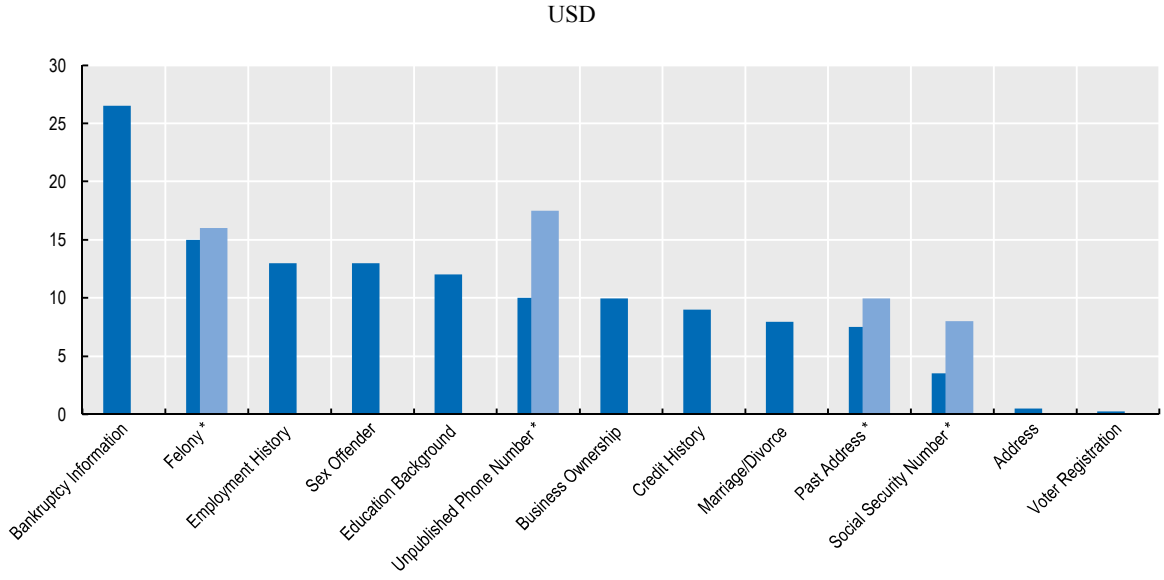
The core business objective of data brokers is to collect and aggregate data, including personal data (FTC, 2014). Data brokers such as Bloomberg, Nielsen, STATS (sports data) and World Weather Online tap into a variety of data sources that are used for data-related services. These include, for example, data that are disclosed or provided by individual firms and citizens; data from firms that install sensors; data crawled from the Internet; and data from non-profit and public sector agencies (e.g. earth observation data and demographic, health and other statistics). Some data brokers also analyse their data sets to provide information and intelligence services to their clients in wide range of domains for a variety of purposes, including verifying an individual's identity, product marketing, and fraud detection. This is where the boundary between data brokers and data-driven entrepreneurs may be blurring (see the section below entitled "Combining internal and external data sets").

Most data brokers focus on the B2B market segment. Businesses can for instance purchase the email addresses of potential customers from data brokers for marketing purposes. In addition, analytical products sold by data brokers can, for example, provide insights on the media channel to be used for advertisement (e.g., online or newspapers) and/or the geographic region to be targeted (FTC, 2014). Other data brokers provide "people search" websites through which users can search for publicly available information about potential consumers extracted from social media or other online content; this allows them to find old friends or obtain court records or other information about consumers or job applicants (FTC, 2014). But the activities of data brokers are not limited to the commercialisation of personal data. Data brokers such as Bloomberg offer, for instance, professional data services based on financial data to businesses. World Weather Online as another example sells global weather forecast and weather content to websites, businesses and the travel industry.

Analysis of the data available on B2B data brokers suggests that they primarily use the following key revenue models: (i) pay-per-use, (ii) licensing, and (iii) freemium (with a subscription-based premium) model. Data brokers that sell personal data for example for advertisement purposes mostly use the *pay-per-data-set* model. Figure 2.3 summarises some estimates that are derived from various online data brokers.[15] These estimates provide some insight into the relative market values of different pieces of personal data (OECD, 2013a). Another example of the *pay-per data set* model is World Weather Online, which provides weather data via its application programming interface (API). The total amount paid by customers is based mainly on their requests per day. Other data brokers such as Nielsen are selling their data sets (and specific analysis results) via their online stores. The pricing scheme is not completely transparent but it seems that the more specific the insights provided are, the higher the price.

For online advertising, revenues are also generated based on auctions. Data exchanges, for example, are marketplaces where advertisers bid for access to personal data about customers. Tracking the Internet activity of consumers is essential for this business model. Within seconds of visiting a website affiliated with a tracking company, detailed data on a web surfer's activity may be auctioned on a data exchange, such as that run by BlueKai (Angwin, 2010; cited in OECD, 2013a). According to their website, the BlueKai Exchange is the world's largest data marketplace, with data on more than 300 million users offering more than 30 000 data attributes; it processes more than 750 million data events and transacts over 75 million auctions for personal information a day (OECD, 2013a). The freemium revenue model is commonly used by data brokers in combinations with other revenue models. The sports data provider STATS, for example, offers a premium service based on monthly subscription to different applications in various price categories. World Weather Online is another example where a freemium model is used in addition to pay-per-data-set model.

Figure 2.3. **Market prices per record for personal data by type, 2011**



* Two different prices provided by different providers.

Sources: Locate Plus (address, unpublished phone number, felony); Pallorium (address, past address, unpublished phone number, social security number); KnowX via Swipe Toolkit (past address, marriage/divorce, bankruptcy information, business ownership); LexisNexis via Swipe Toolkit (education background, employment history, social security number, felony, sex offender); Experian (credit history); and Voters online.com (voter registration).

But data brokers are not limited to collecting and processing data for the B2B market. Some data brokers are also focusing on the business to consumer (B2C) segment, providing consumers with insights on consumer goods and services (Hartmann, et al., 2014). Examples of consumer-oriented data brokers include AVUXI and CO Everywhere; these provide data on local businesses to consumers, including data on restaurants and bars, based on a variety of data collected from the web. The revenue models of these consumer-oriented businesses are primarily advertisement and/or commission fees, which are obtained when successfully providing businesses with customers. Given the high importance of value-added services provided by many B2C data brokers, making the distinction between B2C data brokers and data-driven entrepreneurs – in particular data explorers – is often difficult.

Businesses and consumers can benefit from the services of data brokers, but at the same time are exposed to many risk factors due to the often sensitive nature of the data collected, analysed and provided. While the service of data brokers may help to prevent fraud, improve product offerings and deliver tailored advertisements to consumers, there can be significant negative side effects arising from (e.g.) misguidance of consumers, discrimination, and violation of consumer privacy (see Chapter 5 of this volume). For example, the scoring processes used in some marketing products are not transparent to consumers, rendering them incapable of preventing possible negative effects of data analytics (see FTC, 2014). There may also be a lack of transparency in the revenue schemes. Different clients may have to pay different prices, depending on the type of client (e.g. researcher, firm or government), the size of the client, the markets in which the client is active, and the purpose for which the data are expected to be used. It is equally important to acknowledge that data brokers also provide data to other data brokers, which develop new combinations of data products.

Public sector

Governments are important actors in the data ecosystem; their multiple roles can include data provision and use, investment, and provision of legal frameworks and regulation. As Chapter 10 of this volume highlights, the public sector is one of the economy's most data-intensive sectors. In the United States, for example, public sector agencies stored on average 1.3 petabytes (millions of gigabytes) of data in 2011,[16] making them the country's fifth most data-intensive sector (OECD, 2013b). The public sector is not only a key user of data and data analytics, but also a major *source* of data. However, the circumstances under which the public sector should provide value-added products from its data assets continue to be debated (see Chapter 10).

The public sector has nevertheless led the way in opening up its data to the wider economy through various "open data" initiatives (Ubaldi, 2013) and thanks to government initiatives promoting improved access to and reuse of public sector data (PSI).[17] The OECD (2008) *Council Recommendation for Enhanced Access and More Effective Use of Public Sector Information* (PSI), which is currently under review, describes a set of principles and guidelines for access to and use of PSI including public sector data (see Annex of Chapter 10).

As highlighted in the OECD (2008) PSI Recommendation, public sector data should be provided free of charge. When data are not provided free of charge, pricing should be transparent and consistent across different organisations and not exceed marginal costs of maintenance and distribution to ensure reuse and competition. Evidence shows that reduced pricing (e.g. allowing non-commercial reuse at zero cost and reducing the charges for commercial use) significantly increases the use of open government data (Capgemini Consulting, 2013), and cross-country research underlines the firm-level benefits from free or marginal cost pricing (Koski, 2011). Potential revenue models for the public sector are therefore variants of the freemium model where higher value-added data product or services are sometimes provided in addition to, and for cross subsidising, the free basic data product, if at all. EC (2013) also discusses alternative sources of revenue, including public funding, usage fees and advertisement. Chapter 10 presents an in-depth analysis of open government and PSI initiatives.

Individuals (consumers)

Even individual end users and consumers can become (active) data providers. A number of start-ups, such as Personal, are currently offering so called "data lockers" where people can gather, store and manage their personal data. These services allow people to take control of their personal data and "re-use it to their own benefit" (The Economist, 2012). Another possible development that could prove interesting is the rise of personal data marketplaces. Start-ups like Handshake and Enliken offer platforms where users can sell their personal data to interested parties (Lomas, 2013), and there are many alternative and more open initiatives of consumer participation through crowdsourcing as well. The social traffic app Waze, acquired by Google in June 2013, collects and aggregates data generated by its users to create real-time traffic information. In the Netherlands over 10 000 iPhone owners joined the collaborative research project iSPEX to measure aerosols via their mobile phones.

When analysing business models related to consumers it can be noted that consumers for the most part do not profit in direct monetary terms when providing their data to companies. As discussed before, they may instead profit from "free services" in exchange for their personal data, and they may also be confronted with advertisement. Nevertheless, there are a number of interesting developments with companies such as Handshake that are promoting a marketplace for personal data in which consumers are immediately rewarded financially when providing their personal data. The increasing demand of consumers to gain more control over their own data is also reflected by initiatives such as Diaspora* (OECD, 2012a). Initiated by four students in the United States, Diaspora* aims to highlight the significant discrepancy between the relatively low value that social networking sites provide and the privacy that users are required to give up in return (Dwyer, 2010; Suster, 2010). The project provides a decentralised platform that allows users to save their personal data on their own servers (at home or at the web-hosting provider), and thus makes it possible for the users to own and control their information.

However, anecdotal evidence suggests that the share of businesses aiming to empower individuals to play a more active role in the use of their personal data is still very low compared to the share of businesses aiming at exploiting personal data. The main reason for this can be assumed to be the relatively poor potential for profit in the case of privacy-enhancing services. This is illustrated by the case of Buyosphere, a start-up based in Canada. When the company began operations in 2010, its aim was to help individuals take control of their shopping history, while giving them the possibility of organising it, sharing it and tracking how they influence others. Buyosphere's initial business model was based on a consumer-to-business (C2B) communication flow: rather than businesses gathering personal data on users' behaviour and pushing advertising at them, Buyosphere would give consumers the power to share their preferences with the companies they choose directly. Furthermore, Buyosphere would let consumers port their own data so they could use the data for their own purposes. Tara Hunt, the CEO and co-founder, did admit however that running a C2B retail company had a significant downside by saying, "Well, we can't promote what would make us the most money" (O'Dell, 2011). And so in the course of its first year, 2010, the company redefined its business model, and now provides online product search through a combination of social search and intelligent use of data. Only once embarked on this pivotal transformation was the company able to raise additional USD 325 000 in venture capital (VC).

Analytic service providers

The field of data and IT infrastructure would seem to leave limited room for new entrants, as it is dominated by traditional vendors and a few independent Hadoop distribution providers. Yet there is indeed room for an explosion of start-ups that focus on data analytic services (including the development of software applications and visualisation tools based on data analytics).

The services of these start-ups and SMEs sit on top of the foundation layer of IT infrastructures such as database, Hadoop and analytic software solutions. Cloudera's CEO Mike Olson (Harris, 2011a), discussing the future of Cloudera and its products, noted that he sees great potential for specialised companies in this layered construction. These new companies focus on specific analytical or visualisation solutions, targeting specific industries or even specialised tasks within an industry.

There are pragmatic reasons for this, which are inherent in start-ups. Because of their focused approach, these smaller companies can offer value and ease of use that generic tools lack. As shown by Criscuolo, Nicolaou and Salter (2012), new technologies and innovations are often first commercialised through start-up companies as they are not as captured by the *innovator's dilemma* as incumbents (Christensen, 1997). They can instead leverage the advantage of starting without the legacy of an existing business and customer base to experiment and create a rich variety of presumably new business models (Hartmann et al., 2014). Expert interviews by TNO (2013) emphasise the limitations of generic off-the-shelf tools provided by incumbent IT suppliers. According to Clive Longbottom, founder of the analyst house Quocirca, many IT suppliers have a tendency to sell one-size-fits all offerings, whereas these new start-ups try to cater to very specific data needs (Heath, 2012).

As the need for business intelligence becomes more focused on real-time insights rather than historical and periodical information, the demands from the users of data analytics have changed; there is now higher demand for advanced specialised data analytic services (see Chapter 3). In addition, it is becoming increasingly important not only to generate the best actionable output, but also to present it in such a way that it is aligned with the business process that it strives to support in order to establish competitive differentiation (Dumbill, 2011). As discussed in Chapter 3, it is expected that for the next couple of years most of the value of data will be added by advanced analytical techniques, in particular predictive analytics, simulations, scenario development, and advanced data visualisations (Russom, 2011). These are the most important growth areas for the near future that data analytic service companies are now targeting. The generic analytical tools that are often provided by many IT suppliers can be important building blocks, but as the threshold for competitive, differentiating data analytics increases, data analytic applications need to be optimised for the context in which they will be used, and analytic service providers are often positioning themselves as specialised service providers to do that job.

Analysis of existing data-driven business models by Hartmann et al. (2014) suggests that two types of business models characterise analytic service providers, which they refer to as "analytics-as-a-service" and "aggregation-as-a-service". Data analytic service companies provide advanced data aggregation and/or analysis services to their customers, which are primarily businesses. But the main characteristic of data analytic service providers that distinguishes them from (e.g.) data brokers is that their activities are primarily based on data provided by their customers rather than obtained from crawling the web or collected from third parties including other data brokers. Where external data sources are collected and integrated by data analytic service companies, this is done mainly to enhance the results

of prior analysis of their customer's data. Furthermore, data analytic service providers will typically act as subcontractors to the data controller (i.e. data processor), while data brokers typically act as independent data controllers (in the B2B and B2C market).

The revenue model of data analytic service providers is therefore often based on service contracts. But increasingly, Internet start-ups are providing their services – including the analytic results – via APIs and visualisation platforms (Hartmann et al., 2014). These start-up companies can (and therefore do) use alternative revenue models, including in particular subscription and usage based revenue models. For example, the start-up company Welovroi, based in Madrid, Spain, is a marketing company that provides monitoring and analysis tools for data provided by customers via the Internet. Welovroi offers its services on a monthly subscription basis, the amount of which depends on the number of employees of its customers, and the number of web services that the customers use.

Another interesting development in data analytic services is the crowdsourcing of data analysts. These services enable organisations that include businesses and governments, as well as individuals all over the world to post their data and let others compete to produce the best analytic results. Crowdsourcing of data analytic activities can lead to faster results, on unprecedented scales, and with better quality control than any individual or small research group can attain. Given its open, informal structure, crowdsourcing is cross-disciplinary by design. In some cases, even gifted amateurs and people without direct experience with the problem provide valuable insights and solutions.

InnoCentive, one of the first companies to crowd-source in the chemical and biological sciences, today has more than 300 000 registered "solvers"', who stand to gain rewards of between USD 5 000 and USD 1 million if their solution works. Key to the success of InnoCentive's crowd-sourcing has been: i) a carefully defined governance structure designed to protect intellectual property from both the seeker and the solver; ii) reduced barriers to participation, so that the challenge scales quickly; and iii) global reach, increasing the likelihood of solutions coming from very unexpected directions. Another popular example of a start-up providing crowdsourced analytic services is Kaggle, which in November 2011 raised USD 11 million from a number of investors (Rao, 2011).[18] Hal Varian, Google's Chief Economist, described Kaggle as "a way to organise the brainpower of the world's most talented data scientists and make it accessible to organisations of every size" (Rao, 2011). According to Kaggle, more than 200 000 data scientists have registered worldwide, from fields such as computer science, statistics, economics and mathematics. These data scientists are competing for prizes as high as the USD 3 million *Heritage Health Prize* for the most accurate prediction of the patients who are most likely to require a hospital visit within the next year (The Economist, 2011).

While firms such as InnoCentive and Kaggle aim at data analysts that have advanced skills in data analytics, other crowdsourcing platforms are designed in such a way that the data analytic problem is masked and presented to Internet users in a very simplified way, often it takes the form of a game, which when won leads to the solution of the original data analytic problem. Foldit, for example, is a popular online citizen-science initiative, in which individuals are scored on the structure of proteins that they have "folded". The game records the structure and the moves that the players make, and scientists can capture the data that are then used to improve the problem-solving process in every aspect, from the quality of the scientific results to how long people play the introductory levels meant to teach the game.[19]

Another example is Zooniverse, which enables researchers to design crowdsourcing platforms that take their data and present them in a format that will let the crowd help

them to achieve their objectives. Zooniverse has a community of over 850 000 people, who have taken part in more than 20 citizen science projects over the years. These initiatives support a form of "scientific democracy", where data can be shared among and utilised by investigators in public and private sectors, policy makers, and the public. Crowdsourcing platforms for research and health research in particular are discussed in more detail in Chapter 7 and 8.

Data-driven entrepreneurs

Data-driven entrepreneurs are using data analytics to various ends, ranging from cost saving through financial monitoring to revenue growth through new marketing strategies and product development. As a study by Brynjolfsson and McAffee (2012) points out, these goals strongly depend on the maturity of an organisation in terms of its ability to deploy data analytics and related technologies. As companies gain more advanced data analytics experience, the balance between cost saving and revenue growth will shift. Deploying data and analytics for marketing and sales becomes more important, as does, to a lesser extent, product research and strategy development purposes. More disruptive innovations that upend current business practices, or create new ones, require more experience, greater commitment, and a more solid belief in the potential of leveraging data. Still, the use of data and analytics for incremental changes can be a helpful precursor for more radical disruptive DDI, in which (networks of) organisations rethink products, business models or even whole value chains (Lavalle, 2010).

Although some companies have indeed shifted their data-related priorities from cost efficiency to revenue growth to innovating for competitive differentiation, this has not yet resulted in a grand-scale proliferation of more disruptive DDI since most organisations deploy data analytics to enhance their existing business models. However, examples of such kinds of disruptive innovation are increasing in number; they are realised by start-up companies as well as traditional (non-ICT) companies. These companies base their innovative business models on the deployment of applications that use data generated through the Internet including the Internet of Things (IoT – see Chapter 3). They thus build their products (goods and services) on top of existing data, using that data as an input to provide their innovative goods and services. The US-based start-up BrightScope, for example, extracts public data from the Department of Labor and processes it to bring transparency to opaque markets. Through the use of cloud-based software, the company aims to drive better decision-making in the areas of retirement plans and wealth management.

Two interesting examples of traditional (non-ICT) companies are Nike and the Dutch IJkdijk, which have redesigned some of their traditional products as "data products". Nike introduced the online Nike+ platform, the Nike+ sensor that can be clipped on running shoes, an app that tracks runs and more recently the FuelBand, a wristband that tracks activities and calories burned during the day. Although its core value proposition – supporting people to be physically active and healthy – has not changed, Nike is now more and more providing this proposition by using data that enables users to set their goals, track their progress and include social elements. It has also created an API that allows third parties to develop apps based on this data-driven platform. The IJkdijk is the result of a research program in which a dike in the north of the Netherlands was equipped with sensors. The collected data are analysed and visualised to improve dike monitoring and water management. Both examples illustrate the potential of sensor data and M2M for DDI. Another example is autonomous self-driving cars discussed in Chapter 3. The development of autonomous and smart cars is in line with a bigger transition towards

smart cities in which organisations are deploying data and data analytics to realise innovations in a complex and dynamic environment (see Chapter 9).

Building on their own experience and expertise and their accumulated assets including data and analytics, many data-driven entrepreneurs may become data and analytic service providers for others as well. In this case they are not solely consumers of data and analytic products; they also contribute with their data and software development activities for the benefit of other organisations that can reuse the data and the data analytic solutions for very different purposes. The example of Amazon as a big data powerhouse was already given above (Assay, 2013). Walmart, as another example, is developing its own data analytic services via its subsidiary Walmart Labs,[20] which is also actively contributing to the co-development of open source analytics. Another example is John Deere, which is transforming itself from a manufacturer of tractors to a highly advanced business intelligence service provider for farmers. Finally, there are businesses that open up their data; an example is the Dutch energy network service provider Alliander, which recently organised a workshop with partners and stakeholders to explore the potential of open data.

These examples illustrate how even organisations for which data and analytics originally were not part of their primary business model can become actors for different steps of the value creation process in the data ecosystem. This phenomenon has been described by Rao (2013), who wrote an article about non-tech corporations "eating" tech-start-ups as they try to position themselves since datafication is affecting their market:

> *It's no longer Google, Facebook and Yahoo that are competing to acquire the best and the brightest start-ups in Silicon Valley. There are plenty of corporations in retail, health, agriculture, financial services and other industries that are sending their corp-dev talent to scout out possible acquisitions in the Bay Area and beyond (Rao, 2013).*

Based on Hartmann et al. (2014), two major types of data-driven organisations can be distinguished: i) those that provide goods and services based on the collection of data available on the web and via data brokers (i.e. data explorers), and ii) those that provide goods and services to generate data that are used to enhance user experience and to empower additional services (i.e. data-generating platforms).

Data explorers

Data-driven entrepreneurs that act as data explorers are closely related to data brokers in the sense that they collect available data either by crawling the web, tapping into social media sites, or even purchasing data from brokers. However, in contrast to data brokers – that have as a primary business objective the provision of data and/or of value added insights – data explorers have a well-defined business objective that addresses particular business or consumer needs (other than the need for data or insights). And unlike the data-generating platforms discussed below, data explorers do not deploy the means to generate data themselves. An example of a data explorer is Gild, which helps companies recruit software developers by automatically evaluating the software source code these developers have published on open source software sites such as GitHub and Google Code, and their contributions to popular Internet forums on software development such as Stack Overflow. An expertise score is computed to rank a developer's ability to code, while another score, the demand score, assesses how competitive it will be to recruit the candidate (Gild, 2014).

The revenue model of data explorers depends on whether they are targeting the B2B or B2C market. In the case of B2B, their revenue model is similar to that of online analytic service providers: B2B data explorers tend to rely primarily on freemium (with subscription based premium) revenue models. In contrast B2C data explorer tend to rely more on revenue models-based on advertisement and commission fees, but sometime also in combination with freemium and subscription based revenue models (Hartmann, et al., 2014). For example, DealAngel, founded in 2010 in Moscow, Russia, provides consumers with a list of hotels with the best deals free of charge. Its revenues are generated based on commission fees from the booking websites that consumers are directed to when actually booking a hotel (Ha, 2012; Hartmann et al., 2014).

Data generating platforms

Data generating platforms include a wide range of companies ranging from small, low-tech SMEs to highly data-intensive companies such as Apple and Google, including traditional (non-ICT) companies such as Nike and TomTom. They typically include data-driven service providers (i.e. service platforms) from which data are generated as a by-product of their actual business activity to support the sales of goods and services: this contrasts with data explorers or data brokers, for which the reuse of existing data is at the core of their business models. The service platform providers also include businesses that sell mobile applications (apps) or sensor equipped smart devices that are interconnected via machine-to-machine communication (M2M) in the IoT (see Chapter 3). Companies such as Monsanto, John Deere and DuPont Pioneer are, for example, taking advantage of the "Industrial Internet" by integrating sensors with their latest equipment "to help farmers manage their fleet and to decrease downtime of their tractors as well as save on fuel" (Big Data Startups, 2013). The same sensor data are then linked with historical and real-time data on e.g. weather prediction, soil conditions, fertiliser usage and crop features to optimise and predict agricultural production. In the case of John Deere, some of the data and analysis results are presented to farmers via the MyJohnDeere.com platform (and its related apps) to empower farmers to optimise the selection of crops, and of where and when to plant and plough the crops (Big Data Startups, 2013).

The fact that service platform providers produce data as a by-product does not prevent them selling their data to third parties. For instance, service platforms may share their data with business partners, or may provide a platform (online) that allows the exchange of several information services to clients in a range of domains. Data collected on agriculture platforms such as provided by Monsanto, John Deere and DuPont Pioneer, for example, are being considered as an important data source for biotech companies to optimise genetically modified crops (GMC). Reuse of the data is also being considered by crop insurance companies and traders on commodity markets, which has led to controversial discussions on the potential harm to farmers from discrimination and financial exploitation (Bunge, 2014; *The Economist*, 2014).

The main characteristic of service platforms is that they benefit from data enabling multi-sided markets, where activities on one side of the market go hand in hand with the collection of data, which is exploited and used on the other side of the market (see Chapter 4 of this volume). These markets are also taking advantage of network effects emerging on at least one side. For data explorers and data brokers, in contrast, the characteristics of multi-sided markets are less applicable, but economies of scale, in particular due to network effects, are more relevant. The revenue model of data generating platforms therefore relies heavily on the combination of network effects that typically affect all sides of the market of the service platform provider. As the utility for

users on all sides of the market increases with the increase in their numbers, users are more willing to pay for access to a bigger network and/or to contribute with their own data. Combined with the increasing returns to scale and scope the data enable, these network effects can lead to huge profit margins for platform providers (see Chapter 4).

Cross subsidies are therefore often used by service platform providers: a service is offered for free or at a low price on one side of the market (often the B2C market), but subsidised with revenues generated on the other sides of the market (often the B2B market) (Bonina, 2013). For instance, service platform providers often use the *freemium model* on one or more sides of their market, where the cost of the free service is subsidised by premium customers across all sides of their markets. Online dating portals, for example, operate with a freemium and premium subscription based model on both sides of their market. But often more complex revenue models are used. For example, the freemium revenue model can be used on one side of the market (e.g. the consumer market), sometimes in combination with an advertisement revenue model. In the case where a physical device is required (e.g. navigation system hardware such as provided by TomTom), an asset sale model may be used instead or in addition (Hartmann et al., 2014). For the other side of the market, service platform providers can use the same models as described above for data brokers, namely freemium (with subscription-based premium) model or service contracts. Platforms such as Facebook are financed by (e.g.) advertisers on the side of their market that uses data provided by individuals on the other side of the market. Advertisers can thereby better target potential consumers to increase sales, and individuals have access to social network services that are provided to them free of charge in exchange for the free use of their personal information by Facebook.

2.2. Interactions in the data ecosystem

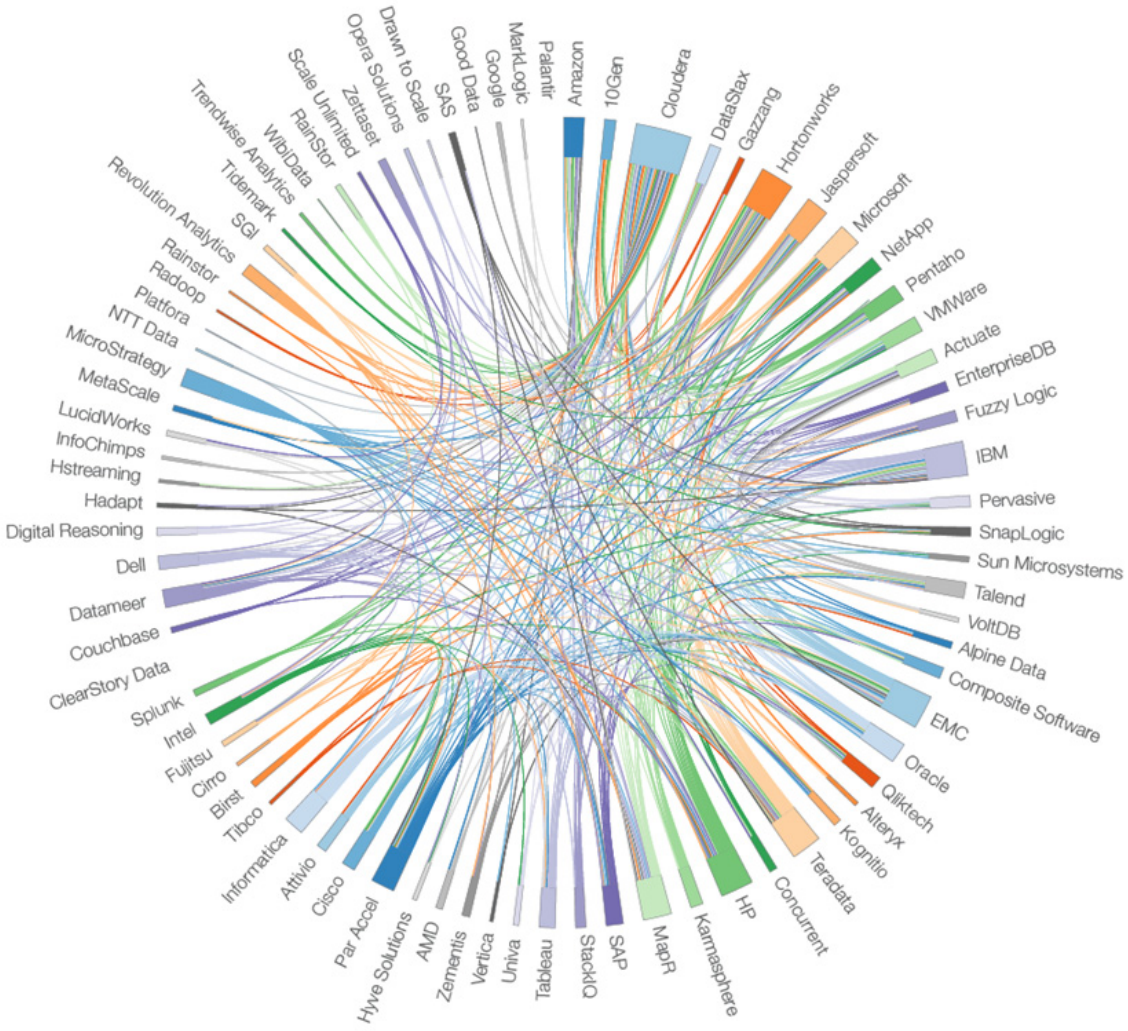
This section discusses the interaction among actors that structure the data ecosystem. It analyses in particular the relation between competition and collaboration for DDI, and how this translates into horizontal and vertical movements by the various actors.

Co-opetition: Competition and collaboration

Competition, but also collaboration, or "co-opetition" is key to leveraging the potential of the multidisciplinary field of DDI (see Woo, 2013). The multidisciplinary characteristics of "big data" challenges and opportunities are not confined to the complex stack of infrastructural elements, analytical techniques and visualisation tools. They also include organisational and HR-expertise and extensive domain knowledge in the specific areas where data are applied. In that respect, the functioning of the data ecosystem fits the general description of innovation ecosystems, in which collaboration among individual companies allows them to create value that no single company can deliver on its own (Adner, 2006).

It is difficult to properly assess what companies are currently dominating the data ecosystem. Exact numbers of market share are hard to come by and would be difficult to interpret as the ecosystem comprises many different kinds of intertwined services. However, if collaboration is a necessity in the data ecosystem, the number of partnerships could serve as a potential indicator of market activities. Figure 2.4 provides an overview of more than 50 companies with the highest number of partnerships with other organisations in the Hadoop ecosystem (O'Brien, 2013).

Figure 2.4. **Partnerships in the Hadoop ecosystem, January 2013**



Note: The larger the source bar, the greater will be the number of a company's partnerships. For example, Cloudera has by far the highest number of partnerships, followed by Hortonworks, IBM, and EMC.

Sources: O'Brien, 2013, based on Datameer, 2013.

The Hadoop ecosystem includes many actors from the different layers presented in Figure 2.2. The IT infrastructure providers Cloudera, Hortonworks and MapR, which act as independent Hadoop distribution providers, are especially well connected, but so are more traditional IT vendors such as IBM, EMC, HP, Microsoft, Oracle, SAP, VMware, Cisco and Intel. Data-driven entrepreneur Amazon and IT infrastructure provider Dell – which did not make the list in 2012 – have both improved their networks considerably in 2013 and are now among the top contributors to the Hadoop ecosystem. As noted before, many actors within the global data ecosystem are active in different layers, or steps in the value creation process. This includes in particular the biggest actors such as Microsoft, Google, Amazon, Oracle, SAP, SAS and VMware.

Looking at the biggest actors in more detail including their economic performance (Tables 2.1-2.3), it is worth highlighting a number of findings:

1. The large providers participating in the Hadoop ecosystem, are mainly companies registered in the United States with the exception of Yahoo Japan, NTT Data, and Fujitsu (Japan), SAP (Germany), Persistent System (India) and Acer (Chinese Taipei). That said, it should be also noted that the number of top providers has been reduced due to merger and acquisitions (M&A).
2. It comes as no surprise that most of the top firms participating as providers in the data ecosystem are Internet and software firm. However some hardware firms and in particular IT equipment firms are heavily involved as well. Semiconductor firm Intel and AMD (Advanced Micro Devises) are the exceptions. In terms of M&A these companies have been remarkably active, which explains their increasing involvement in the data ecosystem.

Table 2.1. **Performance of the top Internet firms involved in the Hadoop ecosystem, 2013**

USD million (except employment numbers)

Internet firms	Country of registration	Revenue 2013	Employment 2013	R&D 2013	Income 2013
Amazon.com Inc	United States	74 452	117 300	6 565	274
Google Inc	United States	56 168	53 861	5 467	14 655
Facebook	United States	7 872	6 818	1 415	1 500
Yahoo! Inc	United States	4 987	11 700	886	3 946
Netflix Inc	United States	4 375	2 045	379	112
Yahoo Japan Corp	Japan	3 795	5 780		1 229
Concurrent Computer Corp	United States	63	229		4

Source: OECD Information Technology database, compiled from annual reports, SEC filings and market financials, July 2014.

Table 2.2. **Performance of the top ICT service and software firms involved in the Hadoop ecosystem, 2013**

USD million (except employment numbers)

ICT service and software firms	Country of registration	Revenue 2013	Employment 2013	R&D 2013	Income 2013
International Business Machines Corp	United States	99 751	434 246	6 226	16 483
Microsoft Corp	United States	77 849	99 000	10 411	21 863
Oracle Corp	United States	37 920	120 000	5 149	10 806
SAP AG	Germany	22 858	66 061	3 102	4 521
Computer Sciences Corp	United States	13 544	87 000		1 501
SYNNEX Corp	United States	10 845	12 500		152
VMware Inc	United States	5 207	14 300	1 082	1 014
Teradata Corp	United States	2 743	10 200	179	395
Informatica Corporation	United States	948	3 234	166	86
Microstrategy	United States	576	3 221	98	83
Splunk Inc	United States	303	1 000	76	- 79
Persistent System	India	284	6 970		42
Tableau Software Inc	United States	232	1 360	61	7
Pervasive Software Inc	United States	49	255		2
NTT Data Intramart Corp	Japan	42	257		1

Source: OECD Information Technology database, compiled from annual reports, SEC filings and market financials, July 2014.

Table 2.3. **Performance of the top ICT hardware firms involved in the Hadoop ecosystem, 2013**

USD million (except employment numbers)

ICT hardware firms	Country of registration	Revenue 2013	Employment 2013	R&D 2013	Income 2013
Hewlett-Packard	United States	112 298	317 500	3 135	5 113
Dell Inc	United States	56 940	108 800	1 072	2 372
Intel Corp	United States	52 708	107 200	10 611	9 620
Cisco Systems Inc	United States	48 607	66 639	5 942	9 983
Fujitsu Ltd	Japan	43 046	168 733		478
EMC Corp	United States	22 787	60 000	2 689	2 557
Acer Incorporated	Chinese Taipei	11 967	7 967	103	- 90
NetApp Inc	United States	6 368	13 060	922	769
Advanced Micro Devices Inc	United States	5 299	10 340	1 201	- 83
Silicon Graphics International Corp	United States	767	1 400	61	- 3

Source: OECD Information Technology database, compiled from annual reports, SEC filings and market financials, July 2014.

Mergers and acquisitions, and vertical integration

As the data ecosystem evolves, many new companies emerge. Subsequently, larger companies try to strengthen their position. Not only will they develop new products and forge partnerships, but they will also acquire promising start-ups to improve and augment their propositions with analytics platforms, visualisations and applications (ESG, 2012). Infochimps CEO Nick Ducoff provides an explanation for this dynamic between the specialised nature of many big data start-ups and the more generic platforms they build on (Watters, 2011b):

> *If you are best at the presentation layer, you don't want to spend your time futzing around with databases […]. What we're seeing is start-ups focusing on pieces of the stack. Over time the big cloud providers will buy these companies to integrate into their stacks. (Watters, 2011b)*

There is a tendency of consolidation in the IT service industry that could also especially affect IT infrastructure providers in the data ecosystem. At the European Data Forum 2013, Siemens manager in charge of the big data initiatives, Gerhard Kress, emphasised the importance of research into vertically integrated algorithms. In an analysis of the big data market ESG,2012, an IT market research and advisory firm noted how data service companies try to obtain dominant positions in certain vertical industries: "[…] where whomever has 'the most data scientists with a vertical bent' may win".

According to a report from Orrick (2012) on emerging big data companies, based on deals and investments mainly in the United States, big data financing activity has increased significantly since 2008 (see Figure 1.4 in Chapter 1 of this volume). Recent years have also seen the take-off of the first IPOs of big data companies. The number of mergers and acquisitions (M&A) has increased rapidly from 55 deals in 2008 to almost 164 deals in 2012, with almost USD 5 billion being invested over that period. In the first half of 2013 alone, big data companies raised already almost USD 1.25 billion across 127 deals. IBM was the most active acquirer of big data companies in 2012, followed by Oracle.

The evolution in value creation with data seems to be reflected in the above described trends on M&A. In the past five years, in terms of both deals and (especially) investments, the focus has shifted from big data infrastructure to big data analytics and applications. Whereas in 2008 infrastructure accounted for 46% of big data investments, this share decreased to 31% in 2012. These numbers also illustrate that the analytics, visualisation and application layer, the "last mile of big data" is where most of the value of data is generated and where true differentiating quality resides as the commoditisation of data analytics continues (ESG, 2012).

Combining internal and external data sets – the emergence of data markets

Most organisations initially apply analytics to their own internal data sets, possibly combining several databases from various departments and processes. But the value of data analytics also lies in the combination of both internal and external data (Redman, 2008). As highlighted in Chapter 4, the value of data is highly context-dependent and "multiplies" when it can be shared and linked with other data sets. As the data are put in a larger context they can reveal additional insights that otherwise would not be possible to glean.[21] A white paper of the European Technology Platform NESSI (2012) stresses how important it is to integrate private data with external data to enhance existing products and services. As O'Reilly's Ed Dumbill (2012a) notes:

> *Mixing external data, such as geographical or social, with your own, can generate revealing insights. […] Your own data can become that much more potent when mixed with other datasets.*

Pointing out that "critical information often resides outside companies", Biesdorf, Court and Willmott (2013) from McKinsey & Company highlight what integrating external data sources involves:

> *Making this information a useful and long-lived asset will often require a large investment in new data capabilities. Plans may highlight a need for the massive reorganisation of data architectures over time: sifting through tangled repositories (separating transactions from analytical reports), creating unambiguous golden-source data, and implementing data-governance standards that systematically maintain accuracy. (Biesdorf, Court and Willmott, 2013)*

In addition to using data from external sources to create value, it could also be valuable to open up proprietary data sets to others. As Rufus Pollock stated at the OECD Technology Foresight Forum in October 2012:[22] "The best thing to do with your data will be thought of by someone else", referring to the open data movement. Chapter 4 highlights a number of reasons why open data can be an optimal strategy from a private and public sector perspective. Some organisations offer their data for free via their website or specific online portals – especially NGOs and governments as highlighted in Chapter 10. Other organisations sell their data. The example of French mobile ISP Orange with its Floating Mobile Data (FMD) technology was already given above. Other well-known examples are Internet firms such Facebook and Google, whose vast collections of personal data are a valuable resource for advertisers. In some cases social media companies work with third parties such as analytic service providers that commercially exploit the social data; examples are Gnip and Datasift. These companies have access to the so-called Twitter Firehose and other social media data, which they prepare and manage to make more accessible and useful to their customers by adding all kinds of filters that fit users' specific needs.

In addition to the data sources and intermediaries mentioned above, including in particular data brokers, data are also exchanged through online services (i.e. data markets) that host data from various publishers and offer the (possibly enhanced) data to interested parties (Dumbill, 2012b). The most established data markets are provided by Infochimp, Datamarket, Factual and Microsoft's Azure, although there are several more (Big Data Startups, 2014). Some data markets try to offer all the data they can, such as Infochimp. Others focus on specific kinds of data, such as Factual, which originally started with location data and is now branching out to a few new specific verticals. Another type of specialisation is to choose a specific target group, such as Figshare, a data market for researchers. The boundaries between data brokers and data market providers are blurred, in particular because both provide the following useful value-added service according to Dumbill (2012b):

1. they provide a point of discoverability and comparison for data, along with indicators of quality and scope
2. they handle the cleaning and formatting of the data, so they are ready for use
3. they provide an economic model for broad access to data that would otherwise prove difficult to either publish or consume.

However, it is interesting to note that despite the growth of data intermediaries, as yet there is no established data marketplace where organisations and individuals can sell or exchange data directly with each other. Some platforms provide some of these functionalities, but they are tailored to specific, tightly integrated value chains that are heavily dependent on each other, for example in mobility, logistics or agriculture (e.g. the "smart dairy project" from TNO and several Dutch companies in the field of dairy farming) (TNO, 2013).

These three propositions illustrate how data marketplaces and data brokers can facilitate finding the right kind of data and fulfil even some additional steps in the value creation process, such as data preparation, to ease further data integration. However, one important distinguishing factor between data brokers and data market providers is that data brokers are actively engaged in the collection of additional data, while data market providers are intermediaries through which data controllers (including data brokers) can offer their data sets. Furthermore, some marketplaces allow their customers to explore data and to mix them together with their own or other available data sets to create new value. Although most marketplaces are focused on developers as their main users, Dumbill (2012b) notes that some data marketplaces try to target less IT-savvy users as well. Microsoft's Azure, for instance, has aligned its data sets not only with its other big data products, but also with its business tools such as Excel. This makes it easier for smaller organisations (and even individual users) to download and combine different (internal and external) data sets. Furthermore, data marketplaces enable a new economic model for data use and sharing, which enhances the overall value of the data provided. As Factual's CEO Gil Elbaz explained at the Strata 2011 conference:

> *Another dimension that is relevant to Factual's current model: data as a currency. Some of our most interesting partnerships are based on an open exchange of information. Partners access our data and also contribute back streams of edits and other bulk data into our ecosystem. (Watters, 2011a)*

The data ecosystem and its global value chains

The data ecosystem involves global value chains (GVCs), formed by companies increasingly dividing up their production processes and locating productive activities in many countries. As highlighted above, the data ecosystem relies on technologies and resources that are distributed around the globe. The ICT infrastructures used to the perform data analytics including the data centres and the software will rarely be located within just one national boarder. They will instead be distributed around the globe to take advantage of factors including local work load, the environment (e.g. temperature and sun light) and labour costs.[23] Data can thus be collected from consumers or devices located in one country through devices and apps developed in another country. They can then be processed in a third country and used to improve marketing to the consumer in the first country and/or to other consumers around the globe.

Furthermore, as highlighted above, many data-driven services stand on the shoulders of giants who have made their innovative services (including their data) available via APIs – many of which are, as noted above, located in foreign countries. One example, which has become better known in developing economies, is Ushahidi, a non-profit software company based in Nairobi, Kenya. Ushahidi develops free and open source software for data collection, visualisation, and interactive mapping based on available APIs provided by Internet firms such as Google and Twitter. One of its first products was created in the aftermath of Kenya's disputed 2007 presidential election to collect eyewitness reports of violence via email and text messages to be visualised on Google Maps. Since then, Ushahidi's data-driven services have been used in particular during crises around the world – for example, in aftermath of the 2010 earthquake in Haiti and the 2010 earthquake in Chile, respectively, where it was used to locate the wounded.

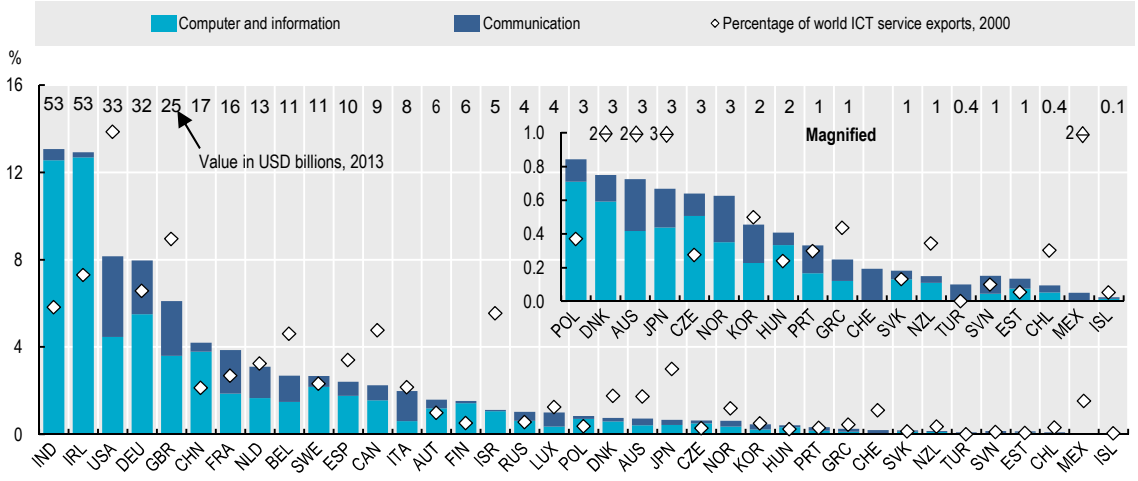
Figures on the distribution of data-driven services are not known. However, the distribution of Internet sites hosted by country code top-level domains (ccTLDs) as identified in the Alexa one million (a list of the top 1 million sites of the world) can provide an approximating picture of how data-driven services are distributed or actually concentrated in the world (OECD, 2014c).[24] The United States alone accounted for almost 60% of all top sites hosted in the OECD area in 2013, or more than 50% of all top sites hosted in the OECD area plus Brazil, People's Republic of China (hereafter 'China'), Colombia, Egypt, India, Indonesia, Russia, and South Africa taken together (see Figure 3.5 in Chapter 3 of this volume). Looking at all sites around the world and grouping the European and Asian countries into regions may give a better perspective. In 2013, the United States accounted for 42% of all top sites hosted, while Europe hosted 31% of the world's top sites and Asia 11% (Pingdom, 2012; 2013). The high concentration of top sites hosted in the United States, which is also reflected in the high number of co-location data centres there (see Figure **1.5** in Chapter 1), are most likely related to a backhaul market that function well in the United States (see Chapter 3). It also supports findings of the US digital data industry's relatively higher share of value-added highlighted in Box 2.1.

As the data ecosystem involves GVCs, many of its activities are captured in international trade. These include not only the trade in ICT services provided by actors in the IT infrastructure layer, but also trade in data-intensive services. As highlighted in Kommerskollegium (2014), even trade involving goods and services that are not data-intensive also typically involve data such as:

- corporate data (to coordinate among different parts of a company and to sell goods and services)
- end-customer data (B2C) (to sell goods and services, enable outsourcing, and provide [24/7] support, and for developing new products)
- human resources data (to co-ordinate among different parts of a company and to match skills, but also to enable outsourcing)
- merchant data (B2B) (to sell goods and services and provide [24/7] support, and for developing new products)
- technical data (to sell goods and services, upgrade software, monitor the operation of products, enable outsourcing and provide [24/7] support, and for developing new products).

There are no figures on data- and analytic-specific services. But taking as a proxy trends in trade in ICT related services, which obviously involve the exchange of data, one can assign a significant growth in cross-border (trade-related) data to the major exporters of ICT services between 2000 and 2012 (Figure 2.5). The largest exporters of ICT services in 2013 were India, Ireland, United States, Germany, the United Kingdom and China. These countries are estimated to be the largest destination of cross-border data. As a consequence, the leading OECD importers of ICT-related services are also the major sources of trade-related data, including in particular the United States and Germany.

Figure 2.5. **OECD and major exporters of ICT services, 2000 and 2013**



Source: OECD (2014d), Measuring the Digital Economy: A New Perspective, OECD Publishing, Paris, based on UNCTAD, UNCTADstat, June 2013, http://dx.doi.org/10.1787/888933148882.

2.3. Key challenges in the global data ecosystem

The globally distributed nature of the data ecosystem, its hyper interconnectedness, and the interdependencies of its actors and their technologies and resources raise a number of policy issues that are specific to the global data ecosystem. These challenges include: i) the difficulty of value attribution which challenges measurement but also taxation policies, ii) the exploitation of key points of control and the competition

implications, iii) the potential barriers to the free flow of data and the importance of the open Internet, and iv) interoperability and standard issues.

Attribution of value, and taxation[25]

The global distribution and interconnectedness of the data ecosystem makes it challenging to attribute the share of the overall value created to specific actors. This has implications for measurement (see Box 2.1), but also raises policy challenges related to taxation. In particular, some governments have expressed concerns that some of the characteristics of the global data ecosystem could create opportunities for *Base Erosion and Profit Shifting* (BEPS) through "aggressive tax planning by multinational enterprises making use of gaps in the interaction of different tax systems to artificially reduce taxable income or shift profits to low-tax jurisdictions in which little or no economic activity is performed" (OECD, 2014e). OECD work on *Addressing the Tax Challenges of the Digital Economy* (2014e) highlights a number of tax issues that the digital economy raises. Many of the issues discussed, however, are not necessarily specific to the global data ecosystem, such as business practices that take advantage of the cross border nature of the Internet to eliminate or reduce tax in a country or that exploit opportunities for BEPS with respect to VAT through e.g. the use of remote digital supplies to exempt businesses (OECD, 2014e).

This section briefly highlights potential BEPS issues discussed in OECD (2014e) that *are* specific to the data ecosystem. Many of these issues emerge due to the global distribution and interconnectedness of the data ecosystem in combination with the economic properties of data discussed in Chapter 4 of this volume. That combination raises a number of questions:

- whether data is being appropriately characterised and valued in corporate balance sheets for tax purposes
- whether any profits attributable to the remote gathering of data by an enterprise should be taxable in the State from which the data is gathered
- and whether current nexus rules continue to be appropriate.

At the core of the issues raised by these questions stands the challenge of attributing the value created in the data ecosystem to specific actors. Attribution is key for the current paradigm used by tax authorities to determine where tax-relevant economic activities are carried out and where value is created. The data ecosystem may challenge this paradigm – and with that, the foundation for taxation in most countries.

Measuring the monetary value of data

The value attribution challenge is most of all related to the challenge of measuring the monetary value of data (see Chapter 4). Most businesses still do not fully take into account the economic value of the data they control in their balance sheet, "although data purchased from another related or unrelated business would be treated as an asset in the hands of the buyer" (OECD, 2014e). As highlighted in Chapter 4, data can under some circumstances be considered a capital good (subject to depreciation). However, in many cases the context dependency of data challenges the applicability of market-based value attribution, since this assumes that markets can converge towards a price at which demand and offer meet. That is not always the case. As "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value" (OECD, 2013a) showed, the monetary valuation of the same data set can diverge significantly

among market participants.[26] Furthermore, where data are collected or generated and no market exists to set a price, businesses may have no means to objectively evaluate their data assets. In that particular context questions have emerged as to whether services provided in exchange for (personal) data can be considered free goods or barter transactions, and how they should be treated for accounting and tax purposes (OECD, 2014e).

Data ownership

Another factor making the attribution of value creation difficult is the challenge related to "data ownership", a concept has turned out to be impractical in many cases (see Chapter 4). In contrast to other intangible assets, data typically involve complex assignment of different rights across different data stakeholders, requiring "the ability to access, create, modify, package, derive benefit from, sell or remove data, but also the right to assign these access privileges to others" (Loshin, 2002). So in many cases no single data stakeholder will have exclusive rights and no clear ownership can be assigned. Different stakeholders will typically have different degrees of rights depending on their role. In cases where the data are considered "personal" the situation is more complex, as privacy regimes typically tend to strengthen control rights of the individuals (see for example the Individual Participation Principle of the OECD [2013c] *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*).[27]

Global distribution and interconnectedness

The distribution and interconnectedness of data-driven services only increase the difficulty of attributing value and ascertaining ownership. As highlighted above, organisations are using analytics not only for their own internal data sets, but increasingly also for combinations of these and external data sets. As OECD,2014e, acknowledges, the attribution challenge "may be exacerbated by the fact that in practice a range of data may be gathered from different sources and for different purposes by businesses and combined in various ways to create value, making tracing the source of data challenging". And the re-combination of resource is not limited to "data mashups[28]". The development of data-driven services based on existing services provided via APIs is another common practice in the data ecosystem. The example of Ushahidi presented above is a good illustration.

That example also points to the global nature of the data ecosystem, which further increases the difficulty of value attribution because of the cross-border nature of the data flows involved. As highlighted above, the data ecosystem relies on cross-border transactions including data collection, processing and use around the world. Determining the functions (and countries) to which profit should be attributed continues to raise severe tax challenges (OECD, 2014e). Where multi-sided markets are used with the groups of customers from each side of the market spread around the globe, the attribution of profit becomes even more challenging, also given the use of different revenue models including freemium and the importance of cross-subsidies. Some have therefore raised the question of "whether the remote collection of data should give rise to nexus for tax purposes even in the absence of a physical presence" (physical establishment, PE), in which case non-resident enterprises could be taxed based mainly on their domestic activities involving data collection. It should be noted at this point that current tax treaties do not permit the taxation of business profits of non-resident enterprises in the absence of a PE to which these profits are attributable, raising further questions on the feasibility of data-based taxation (OECD, 2014e).

Exploitation of the key points of control, and competition

As highlighted above, actors can play different roles in the data ecosystem. Depending on their role and their market power, they will have more or less direct influence in shaping the ecosystem. The growing number of M&A activities related to big data businesses and the transformation of some businesses towards vertical integration described above are just two of the trends through which the data ecosystem is shaped. In addition, some dominant actors in the data ecosystem may have significant control and power over certain activities shaping the system. This section discusses the main points of control through which the data ecosystem can be influenced and eventually disrupted, focusing primarily on the main layers of the data ecosystem presented in Figure 2.2. The section builds on the control point analysis of the Internet developed by Clark (2012), which is a method for "determining which actors obtain power, economic or otherwise, by virtue of control over key components of the system".[29]

The data ecosystem contains a rich mix of control points that are distributed across the layers illustrated in Figure 2.2 and that may differ significantly across sectors.[30] The exploitation of these control points can raise serious competition and consumer protection concerns, as they can lead to the reduction of consumer choice and anticompetitive behaviour. Good governance of these points of control is therefore essential from a public policy perspective to assure that DDI leads to growth and the well-being of all members of society. Their identification is not, however, a simple task, especially given the rapidly evolving nature of the data ecosystem highlighted above.

Clark (2012) presents two criteria that can be used to identify points of control. These criteria assess the degree to which actors in the Internet ecosystem are controlled by others. They are:

- *The degree of choice*, which essentially tests whether users can "route around" a misbehaving actor. Where users have few possibilities to escape the control of the other actor, or where an actor has control over the choices of the users, a strong point of control can be assumed. A minimum level of choice is necessary for competition but also for trust in the data ecosystem as the ability to select among actors allows choosing those that are trustworthy. As Clark (2012) explains in the context of the Internet:

  If the user is to 'route around' a misbehaving actor, the design of the system must give the user that degree of choice. The tussle of control is often thus a tussle over who controls the choice. Examples … include which ISP to use, which DNS to use, which browser to use, and there are more subtle and complex choices that are embedded in the control picture. (Clark, 2012)

- *The degree of confidentiality and privacy*, which essentially tests "what options the actor has to observe what is being done" (Clark, 2012). The capacity to monitor and profile users, including in particular through the collection and analysis of personal data, creates a (soft) power of influence that enables actors to influence and perhaps limit the choices of users. As the monitoring and profiling activity becomes more exclusive, that power will grow commensurately.

Where users have few possibilities to escape the control of an actor, or where users have no real possibility to escape from the observation of an actor, a strong point of control can be assumed, at least at first. Alternatively, increasing the degree of choice and the confidentiality and privacy of users can mitigate the potential misuse of points of control in the data ecosystem. Possible measures to increase choice include those that

enhance interoperability through open standards, and data portability (see further below).[31] Possible measures to enhance privacy and confidentiality include privacy enhancing technologies including cryptography and privacy regulation (see Chapter 5).

The key points of control discussed in the sub-sections that follow focus on the main layers of the data ecosystem. While the Internet provides "a range of design principles that different actors use to 'blunt the instruments of control' by other actors" (Clark, 2012) (e.g. multi-homing, user-selected routes), this facility is less available in the case of the data ecosystem: lock-in and lack of interoperability are still common in some layers, notably in the IT infrastructure layer and the entrepreneur layer. In these layers users may have greatly narrowed choice once engaged with an actor, suggesting that these two layers are likely the strongest points of control. In addition, the layered structure of the data ecosystem, in which actors rely on services provided by the underlying layers, suggests that the power of control may be asymmetric in favour of the actors *in* the underlying layers. In that respect, ISPs have the strongest potential influence on the data ecosystem, as "they exercise ultimate control: if they do not forward packets, the operation fails" (Clark, 2012).

Internet access[32]

ISPs are the regional gatekeepers that provide access to the Internet through the physical transport infrastructure. While some ISPs are going further up the value chain of the data ecosystem by providing IT infrastructure, data and analytic services, most, if not all, still rely on their traditional business models, which consist of granting subscribers (businesses and consumers) access to the Internet. Having realised that they have "ultimate control" on the data flows on which the data ecosystem relies, some ISPs are looking into means for taking advantage of their position to generate more revenues, for example by, differentiating between classes of Internet traffic (e.g. gold, silver bronze) or by dedicating broadband capacity to certain applications including real-time applications requiring timely data transmission and guaranteed delivering times (i.e. quality of service, see OECD, 2014a).

The reorientation of ISPs' business models towards traffic prioritisation and discrimination of applications has raised a number of concerns among other actors in the data ecosystem. These concerns, which some have framed using the term "net neutrality"[33] are not however specific to the data ecosystem. The same concerns have been raised for example in the context of smart applications, such as connected television that expand and place additional capacity demands on the Internet (OECD, 2014a). That said, it is also true that as DDI becomes a new source of growth, the control of data flows become more and more critical.

Some have suggested that traffic prioritisation and the discrimination of applications could transform the business model of ISPs into a two-sided market (OECD, 2014a). In such a market, ISPs could impose charges on content or application providers in addition to end users. However, as OECD, 2014a, highlights, existing offers by ISPs to ensure fast delivery of content have not sufficiently attracted content providers so far. There are several reasons for this, one being that content providers have incurred other costs in order to improve the quality of their service, principally by building or contracting for Content Distribution Networks (CDNs). By caching content at multiple sites, the content provider can shorten the path that content must travel to reach an end user, thus increasing quality and reducing the resources needed for transport of the content over the Internet. As OECD,2014a, highlights, CDNs may be a way to balance the concerns of

policy makers that, on the one hand, content providers should have tools available to increase the quality of their service, but on the other investment in new applications should be encouraged so as not to put new content providers at a disadvantage relative to incumbents with respect to delivery of their application over the Internet.

There is one other aspect of the relationship of ISPs with other actors of the data ecosystem, including content/application providers. As stressed by OECD,2014a, to the extent that an ISP is vertically integrated into content/application provision – and so moves higher up the value chain, as mentioned above – it is important to remain alert to the possibility that it may have the incentive as well as the ability to behave anticompetitively with respect to independent content/application providers.

Application programming interfaces

Looking more closely at the IT infrastructure layer, proprietary solutions (including APIs) are strong potential points of control that could be exploited through vendor lock-in[34] and other anticompetitive measures. In the case of cloud computing, for example, recent surveys among potential cloud users have highlighted a lack of standards and of widespread adoption of existing open standards as one of the biggest barriers to the use of cloud computing (OECD, 2014b). Fear of potential vendor lock-in is often indicated as the reason. The lack of open standards is a key problem especially when it comes to the model of "platform as a service" (PaaS). In this service model, APIs are generally proprietary. Applications developed for one platform typically cannot easily be migrated to another cloud host. While data or infrastructure components that enable cloud computing (e.g. virtual machines) can currently be ported from selected providers to other providers, the process requires an interim step of manually moving the data, software and components to a non-cloud platform and/or conversion from one proprietary format to another. Consequently, once an organisation has chosen a PaaS cloud provider, it is – at least at the current stage – locked in (OECD, 2014b). Some customers have raised the concern that it will be difficult to extract data from particular cloud services that prevent some companies or government agencies from moving to the cloud. Another concern linked to this is that users can become extremely vulnerable to providers' price increases. This is the more relevant as some IT infrastructure providers may be able to observe and profile their users to apply price discrimination to maximise profit (see Chapter 5 of this volume).[35]

APIs, highlighted in this chapter as the "cytoplasm" lying between the layers of the data ecosystem, could thus be exploited as strategic point of control, for example by limiting users' choice in the applications used on top of a service provided over an API (see the example of Twitter in Box 2.4). Trends towards more closed APIs are therefore raising concerns among some actors that rely on open API for their innovative services. This is particularly relevant in view of the recent debate on the ability for legal entities to copyright APIs. This debate has gained significant momentum after a recent petition by the Electronic Frontier Foundation (EFF, 2014) to the United States Supreme Court in November 2014. The petition follows a court finding earlier in May 2012 that Google had infringed on Oracle's copyright on Java APIs in Android, "but the jury could not agree on whether it constituted fair use" (Duckett, 2014).

Box 2.4. **Competitive effects of Twitter's vertical integration**

Twitter's application programming interface (API) allows outside developers to build apps that can pull in information directly from Twitter to display in their own apps. The availability and openness of proprietary API's have been instrumental for the rapid expansion of apps and the growth of platforms such as Twitter.

Twitter has been pursuing a vertical integration strategy by acquiring and building a portfolio of apps. The company purchased apps such as TweetDeck (2011), Tweetie (2010) and Summize (2008) intending to later transform them into brand extensions that serve different platforms and services, e.g. search engines.

The result of this integration is that Twitter wants developers to start building apps that use Twitter, rather than Twitter apps. Twitter has been discouraging developers from using their APIs to make apps that compete directly with their platform, by rejecting apps that rely on tweet feeds via its API and by revoking API access. The risk of such an approach for Twitter or other growing platforms is that the uncertainty of future access to the API will stifle investment and innovation.

In August 2012, Twitter restricted the number of individual user tokens for an app that could access their APIs to 100 000. This essentially means that app developers are limited to 100 000 app installs on users' devices without special permission from Twitter to increase the number. Some developers were forced to require all members to re-login to free up unused keys for new users.

Source: OECD, 2013d, based on Musil, 2011; Mashable;[36] Twitter, 2012; and Yahoo News, 2013.

Intellectual property rights

The issue of API copyright highlighted above directly points to the role of *intellectual property rights (IPRs)*, which is often used strategically in the IT infrastructure layer as a key point of control (see OECD, 2015). This remains true despite the increasing use of open source software (OSS) applications, which have eased some of the constraints that IT infrastructure users have faced in the past (see Chapter 3). For example, some have expressed concerns that the patent US 7650331 B1 on MapReduce awarded to Google could put at risk companies that rely on the open source implementations of MapReduce such as Hadoop and CouchDB (Chapter 3). Such concerns may be justified, but given that Hadoop is widely used today – including by large companies such as IBM, Oracle and others, as well as by Google – expectations are that Google "obtained the patent for 'defensive' purposes" (Paul, 2010).[37] By granting a licence to (open source) Apache Hadoop under the Apache Contributor License Agreement (CLA), Google has officially eased fears of legal action against the Hadoop and CouchDB projects (Metz, 2010).

Data

Access to data can become a critical point of control in this ecosystem, where value creation and competitive advantage are directly related to the capacity to extract insights from (observed) data. The analysis of data can have a significant impact on confidentiality and the privacy of other actors, to the extent that these actors can be influenced and their choice eventually limited. Chapter 5 of this volume discusses in detail the risk of price discrimination, which is one possible way of exploiting data as a strategic point of control.

As actors across the data ecosystem acquire and control massive (proprietary) data sets, there is an increasing risk that "we're kind of heading toward data as a source of monopoly power", as Tim O'Reilly highlights in an interview with Bruner (2012). The risk of "monopoly power", however, must be assessed carefully on a case-by-case basis,

as it will typically depend on the extent to which data can be exploited as a control point. This in turns depends on factors such as the market (segment) under consideration, in particular its rate of technological change;[38] the data sources used; the degree of detriment to consumer welfare; the potential barriers to entry, including the level of investments required for building comparable data sets; and last but not least, other control points such as APIs and IPRs used sometimes in combination with data. Furthermore, it may also depend on the available means to escape the control of the dominant actor, including in particular the availability of open standards and data portability.

For example, access to points of sale (including to consumers' personal data), which is controlled by a single dominant data-driven enterprise, can become a strategic point of control that, if abused, could raise consumer protection and competition issues. In 2011 the Financial Times (FT), for example, pulled its iPad and iPhone apps from Apple's App Store after several month of negotiation. The primary rational for FT's reaction was not because 30% of revenue had to be shared with Apple, but to "keep control of customer data obtained through subscriptions" (Reuters, 2011). By switching its app to the open standard *HTML5* (see Box 2.5), the FT was finally able to bypass Apple's control, and to directly interact with iPad and iPhone users and so gain access to their data. As a consequence, the FT was able to gain more insights into its customers and increase the number of its digital subscribers by 14% within a year (Miller, 2013).

Box 2.5. **HTML5: An open standard for browsers, apps and operating systems**

HTML5 is an update of the HTML standard that dictates how content is displayed on the web. It will affect three key areas of the app ecosystem: i) mobile browsers, ii) mobile apps, and iii) mobile operating systems.

- *Browsers*: HTML5 is the next iteration of HTML, the web mark-up language that tells browsers how to display web pages. HTML5 is a significant evolution of the standard, in that it introduces richer functionality that allows websites in a browser to mimic the functionality of standalone apps.[39] *Strategy Analytics* (2011) estimates there were 336 million HTML5-capable smartphones sold in 2011, and predicts the number of HTML5-compatible phones sold in 2013 will reach 1 billion. One of the key benefits for app developers using HTML5 in a browser is that they are not tied to an app store that may require a share of app revenues. Despite advancements in the HTML standard, native apps (built specifically for one platform) often can make better use of specific hardware features of phones to deliver content, and often run faster than HTML5 content because they are tailored to a specific device or operating system.
- *Apps*: HTML5 can also be used as the core of standalone apps that can be written once and work across different mobile operating systems. The HTML5 can be viewed directly via a browser or "wrapped" into an app that is specific to a mobile operating system, so that the app that can take full advantage of the hardware potential of devices. These new hybrid solutions are emerging from companies such as PhoneGap and Marmalade. With the open-source PhoneGap, developers can write applications using HTML5, JavaScript and CSS, and then compile native apps using PhoneGap to take advantage of APIs for accelerometers, the camera, compass, etc.
- *Operating systems*: HTML5 content is available across platforms via HTML5-compliant browsers, but the emergence of new browser-based operating systems such as Chrome OS and Firefox OS that run apps could further promote use of the standard. In particular, Firefox OS from the Mozilla Corporation will only run HTML5 apps.

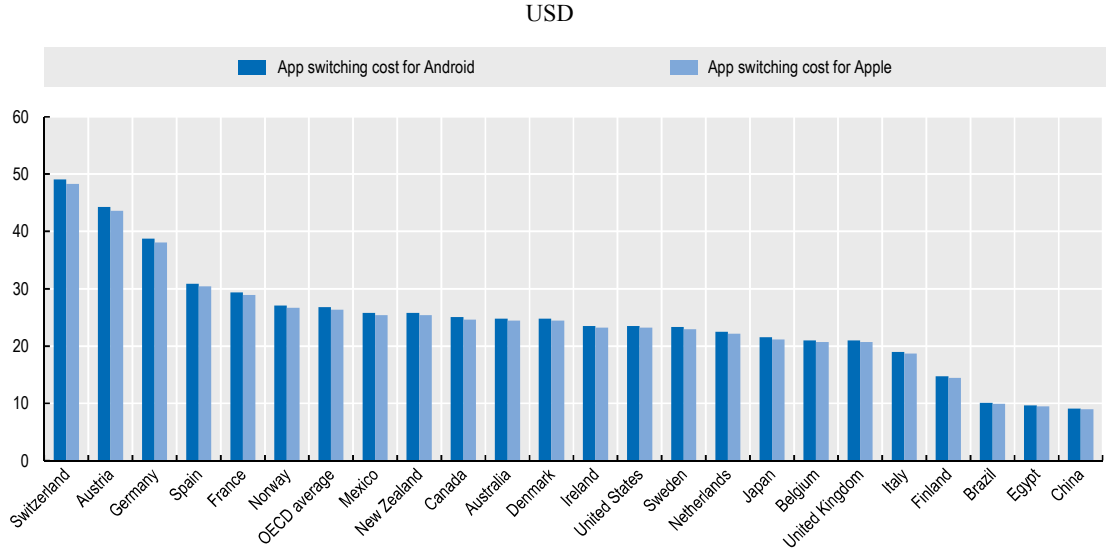
Source: OECD (2013d).

Walled garden

When it comes to multi-sided markets, the situation is more complex: points of control, including data but also other control points such as APIs and IPRs, can be exploited to command multiple sides of the market and thus create a "walled garden" (i.e. closed proprietary platforms). In the case of mobile application (app) platforms (e.g. Apple's App Store and Google Play), for example, consumers may be locked in due to upfront investments in both the hardware and software needed to access the platform, and barriers to data portability that prevent them from reusing their data in other applications. Developers may also be locked in due to upfront investments needed to develop the applications (including in particular the skills and competencies needed). Platform providers can consequently exploit the "stickiness" and lock-in effects of their platforms to reinforce their positions on all sides of their markets. It is important to note that these risks are not restricted to IT services, but this will increasingly also involve physical infrastructures and hardware, as the IoT becomes the dominant network.

Analysis (OECD, 2013d) of the average value of paid apps across countries shows that investments in paid apps could lead to consumer stickiness when it comes to switching platforms (Figure 2.6). Swiss users have an average of nearly USD 50 of paid apps on their phones. The OECD area average is estimated to be roughly USD 26 and these investments would be lost when switching platforms. It is important to highlight though that these sums do not cover all switching costs. They are only a relatively small percentage of the overall purchase price of many smartphones, which will also need to be replaced. Furthermore, as OECD, 2013b highlights, consumer stickiness to a specific platform will also be highly influenced by the amount of digital content that has been purchased on the platform and locked by digital rights management (DRM), as well as the data that have been generated over time.[40] According to Goldman Sachs, the explicit switching cost comes to an average of USD 122 to USD 301 per Apple's iOS device.[41]

Figure 2.6. **App switching costs by platform and by country, 2012**



Note: Data on the average number of apps on Apple's iOS platform were not available so the average number of paid Android apps was used with iOS prices to compute the Apple component.

Source: OECD, 2013d, based on *Think with Google* survey, "Our mobile planet" (2012), www.thinkwithgoogle.com/mobileplanet/en/downloads/.

Competition implications[42]

As highlighted in Chapter 4, the accumulation of data can lead to significant improvements in data-driven services that in turn can attract more users, leading to even more data that can be collected (positive feedback). For example, the more people use services such as Google search, recommendation engines such as provided by Amazon, or a navigation system by TomTom, the better services will be as they become more accurate in delivering requested sites and products, and in providing traffic information. As a result, the service can attract more users. Where data linkage is possible, the diversification of services can lead to further positive feedback. This feedback, which is also characteristic of markets with network effects, finally reinforces the market position of the service provider and has a tendency to lead to its market dominance, or at least to higher market concentration. As Shapiro and Varian (1999) highlighted: "positive feedback makes the strong get stronger and the weak get weaker, leading to extreme outcomes".[43]

Case-by-case analysis of the situation may be required because the degree to which competition issues emerge will typically depend on a number of factors, as highlighted above. However, there are a number of factors that make this analysis particularly difficult, and these may challenge the traditional approach used by competition authorities for assessing potential abuses and harms of market dominance and mergers. The following three types of challenges are highlighted: challenges in i) defining the relevant market, ii) assessing the degree of market power, and iii) assessing potential consumer detriment. As will become clear in the following sections, the factors behind these challenges are basically common to those questioning the attribution of value discussed above.

Challenges in defining the relevant market

Competition authorities rely on a definition of the relevant market as "one of the most fundamental concepts underpinning essentially all competition policy issues, from mergers, through dominance/monopolisation to agreements" (OECD, 2012b). It is the "analytical framework for the ultimate inquiry of whether a particular conduct or transaction is likely to produce anticompetitive effects" (OECD, 2012b). Defining the relevant market is necessary for assessing the effective competition level, including whether an incumbent with significant market power is vulnerable to new competition. Factors in the market definition process will typically include consideration of the goods and services which are perceived by consumers as substitutable, the geographic market, and a time dimension reflecting technological change and changes in consumer behaviour. Given the particular properties of the global data ecosystem, however, establishing a proper market definition can be particularly difficult for the following reason.

Multi-sided markets, such as enabled by data, challenge the traditional market definition, which generally focuses on one side of the market. That approach would tend to define the relevant market too narrowly in a multi-sided market case. As Filistrucchi et al. (2014) argue in the case of two-sided markets: "only in the case of a two-sided non-transactional market, and only when on side does not exert an externality on the other side, can one proceed to define the relevant market on the first side irrespective of the presence of the other side". In many cases however, multi-sided platforms must co-ordinate demand among the interdependent customer groups, and price changes on one side of the market will have "positive feedbacks on the other sides of the market"

(OECD, 2009). In the particular case of data-generating platforms, as has been highlighted, the motivation for the creation of multi-sided markets enabled by data is in many cases founded on exactly these positive feedbacks and externalities that data enable. As a result, focusing on one side of market will rarely lead to a proper market definition. An extended market definition is also justified due to the cross-subsidies often used across multiple sides of the platform. In other words, a proper market definition would have to include all sides involved in the cross-subsidy. Overall, as OECD (2009) highlights, it cannot be assumed that market power and abuse are any less prevalent in multi-sided markets than in traditional markets.

Challenges in assessing market power

Only once the relevant market has been properly defined, can the market power of the market participants be assessed. "[M]arket power can be thought of as the ability […] to sustain prices above competitive levels or restrict output or quality below competitive levels" (OFT, 2004).[44] However, a large share of data-driven products are provided for "free" in exchange for access to personal data, and/or in addition to an offer of a premium version as in the case of the freemium revenue model. In these cases information on prices for the single product will rarely be available, rendering it difficult to assess the degree of market power if applying the narrow market definition discussed above. Other mechanisms therefore have to be used, including in particular a proper market definition: as the data provided will typically be used for different purposes across multi-sided markets, market power will need to be assessed in most cases across all sides of the market as well. As Evans (2011) explains:

> *The existence of a free good signals that there is a companion good, that firms consider both products simultaneously in maximising profit, and that commonly used methods of antitrust analysis, including market definition, probably need to be adjusted to properly analyse two inextricably linked products. (Evans, 2011)*

Nor will assessing market value through the economic value of the collected (personal) data be helpful in most cases; as data have no intrinsic value, as already highlighted above. Admittedly, as possession of that data is necessary for a business to succeed, it can be assumed that the data have economic value. However, the monetary valuation of the same data set can diverge significantly among market participants and uses. This implies that focusing on the ability to sustain prices above competitive levels, is a less practical approach for assessing market power. The restriction of output or quality (to below competitive levels) should be more strongly considered by competition authorities. However, the dimensions that should be included as quality criteria are still not clear – in particular in regard to privacy, which some have argued should be considered when assessing the anticompetitive effects of a particular conduct or transaction (see next section).

Challenges in assessing potential consumer detriment

Anticompetitive behaviour and mergers are often assessed based on the consumer detriment or reduction in consumer welfare they could induce. However, in the particular case where data-driven services rely on personal data, privacy harms are still not fully acknowledged by competition authorities, which will tend to direct the specific privacy issues to the privacy protection authorities; the latter, however, have no authority over competition issues.

The degree to which privacy harms should be considered when assessing anticompetitive behaviour and mergers is therefore still an ongoing debate. That debate was triggered most notably by the former Commissioner Pamela Jones Harbour's dissent in the Federal Trade Commission decision (FTC, 2007) to clear the Google/DoubleClick merger. The dissent was based inter alia on concerns that "the network effects from combining the parties' data would risk depriving consumers of meaningful privacy choices" (Cooper, 2013). Harbour and Koslov (2010) therefore called for competition authorities to consider whether "achieving a dominant market position might change the firm's incentives to compete on privacy dimensions" and thus to promote development of innovative privacy-enhancing technologies and services.

This underscores the need for further dialogue among competition, privacy and consumer protection authorities on potential detriment due to DDI. A preliminary EDPS (2014) opinion confirms that:

> *There is currently little dialogue between policy makers and experts in these fields. […] It is essential that synergies in the enforcement of rules controlling anti-competitive practices, mergers, the marketing of so-called "free" on-line services and the legitimacy of data processing are explored. This will help to enforce competition and consumer rules more effectively and also stimulate the market for privacy-enhancing services. (EDPS, 2014)*

At this point it is important to emphasise that the competition issues discussed above should not be neglected by competition authorities, even when they are engaged in a dialogue with privacy and consumer protection authorities. DDI does not always involve personal data, and the competition issues raised above may still occur in the case of non-personal data; in those cases privacy and consumer protection authorities may have no jurisdiction. The accumulation and control of M2M and sensor data, for example, may raise a number of competition issues in the near future, as data and analytics are increasingly used in areas such as manufacturing and agriculture where non-personal data may become a strategic point of control as well.

Furthermore, it should be highlighted that most competition jurisdictions only enable their authorities to block or challenge anticompetitive practices in which the consequent lessening of competition leads to detriment. If no competition issues are raised, competition authorities will have no jurisdiction. For example, a merger between two companies that do not in any way compete, but whose data sets when combined create links that harm the privacy of consumers, would generate a detriment, but no loss of competition. It that case, competition authorities may not have the right to take action, but consumer and/or privacy protection authorities would.

The free flow of data, and the open Internet

The free flow of information and data is not only a condition for information and knowledge exchange, but also a vital condition for the globally distributed data ecosystem as it enables access to GVCs and markets. As stated already in the *OECD (1985) Declaration on Transborder Data Flows*, "these flows acquire an international dimension, known as Transborder Data Flows", which also favour trade between countries and global competition among actors in the data ecosystem (see Annex of this chapter). In other words, barriers to the free flow of data can limit the effects of DDI by limiting trade and competition, for example.

Some of the barriers to the free flow of data are the intended or unintended results of measures affecting the openness of the Internet (see Chapter 3). These include technical means such as IP package filtering, used *inter alia* to optimise the flow of data for specific purposes, or "data localisation" efforts, either through territorial routing or legal obligations to locate servers in local markets. The social and economic effects of limiting the openness of the Internet are still unknown, although a number of studies have tried to assess the economic costs of barriers. A 2014 working paper by the European Centre for International Political Economy (ECIPE, 2014) aims to quantify the losses resulting from data localisation requirements and related privacy/security laws in seven jurisdictions. According to these estimates, data localisation requirements may result in considerable GDP losses if economy-wide requirements were to be introduced on top of existing privacy/security legislation.[45] However, this study conflates data localisation requirements and privacy and security legal requirements. A more comprehensive analysis is therefore needed that would separate out the economic effects of data localisation requirements from privacy, security and IPR regulations.

There is common interest among countries in finding consensus on how to maintain a vibrant and open Internet and in exchanging views on better practices. The OECD's High-Level Meeting on the Internet Economy on 28-29 June 2011 discussed the openness of the Internet and how best to ensure the continued growth and innovation of the Internet economy. The resulting draft communiqué, which led to the OECD (2011b) *Council Recommendation on Principles for Internet Policy Making*, contains a number of basic principles whose goal is to help ensure that the Internet remains open and dynamic, that it "allows people to give voice to their democratic aspirations, and that any policy-making associated with it must promote openness and be grounded in respect for human rights and the rule of law". The following first five principles are highlighted here as highly relevant for the use of data. This is not to say that other principles are less important to DDI overall:

1. promote and protect the global free flow of information
2. promote the open, distributed and interconnected nature of the Internet
3. promote investment and competition in high speed networks and services
4. promote and enable the cross-border delivery of services
5. encourage multi-stakeholder cooperation in policy development processes.

Interoperability and (open) standards

Barriers to the free flow of data are an issue not only across borders but also across sectors and organisations, including between organisations and individuals (consumers and citizens). Many actors in the data ecosystem still face barriers to data interoperability and portability. Despite the widely agreed benefits, there are still significant (non-legal) issues limiting data exchange and interoperability. This is in particular the case in sectors that require significant investment with a high threshold for new entrants, and more especially capital-intensive industries. The datafication of agriculture, for instance, enables new services from start-ups like Crop-R, but it is driven by innovations from incumbents like John Deere, Monsanto and Lely that enhance their machines and tools with sensors and connectivity to capture and use the data. Interoperability and standards enabling data exchange across the different incumbents' services can be crucial for start-ups like Crop-R.

Interoperability

Open Internet standards such as TCP/IP and HTML5 are crucial for the global data ecosystem – which, as highlighted above, heavily relies on the open Internet for its functioning. In addition, the reuse of data and of data-driven services underlines the importance of (open) standards related to APIs and data formats (including the metadata). However, the lack of appropriate standards that results in potential vendor lock-in and vendors' exploitation of control points in the data ecosystem is still an issue for many users. Vendor lock-in in the cloud computing industry was mentioned above; attempts have been made in that industry to extend general programming models with cloud capabilities in order to enhance interoperability, in particular for PaaS (Schubert et. al., 2010). However, these attempts have not met with success. Promoting open standards for APIs and further work on interoperability are therefore seen as the appropriate response to this problem. As a result many initiatives are under way, covering the full spectrum from infrastructure standards – such as virtualisation formats and open APIs for management – to standards for web applications and services, security, identity management, trust, privacy, and linked data.[46]

But even if data can be extracted, reusability will typically be limited if data are not machine readable and cannot be reused across IT systems (i.e. data interoperability, see Box 2.6). Data are rarely harmonised across sectors or organisations as individual units collects and/or produce their own set of data using different metadata, formats and standards. This means that even if access to data is provided, the data cannot be reused in a different context. This can make it difficult to reuse data for new applications in particular. Unresolved interoperability issues are therefore still high on the e-government agendas of many OECD countries (see Chapter 10). For instance, interoperability of data catalogues, or the creation of a pan-European data catalogue, is a major challenge EU policy makers are facing at the moment.

Box 2.6. **The role of standards for data interoperability**

Reusability of data typically requires that data are machine readable and can be reused across IT systems (i.e. interoperability). Some data formats that are considered machine readable are based on open standards such RDF (Resource Description Framework), XML (eXtensible Markup Language), and more recently JSON (JavaScript Object Notation). But other standards include file formats such as CSV (comma-separated values) and proprietary file formats such as the Microsoft Excel file formats.

To further enable data linkage, meta-data are often needed. They provide the context without which primary data cannot be accessed, linked, or fully understood. As data become abundant and data analytics increasingly automated, finding and making sense of data often requires meta-data. As Cukier (2010) illustrates, meta-data make (primary) data "useable and meaningful as a large library is useless without a card-catalogue system to organise and find the books". Meta-data can be categorised in several types depending on their purpose (see NISO, 2004). Some metadata are provided as open standards, such as the Dublin Core Metadata Term*s*, which defines 15 meta-data elements for describing (web and physical) resources.[1]

1. The Dublin Core Metadata Terms were endorsed in IETF (Internet Engineering Task Force) RFC 5013 and ISO (International Organization for Standardization) Standard 15836-2009.

Data portability

An important development in the area of data portability and interoperability is the increasing role of consumers in the data ecosystem. As highlighted above, consumers play an important role in promoting the free flow of their own personal data across organisations. This role is strengthened by the Individual Participation Principle of the OECD (2013c) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) (see Chapter 5 of this volume). Government initiatives are promoting *data portability* and are thus contributing to the promotion of the free flow of data as well. In 2011, a government-backed initiative called "midata" was launched in the United Kingdom to help individuals access their transaction and consumption data in the energy, finance, telecommunications and retail sectors. Under the programme, businesses are encouraged to provide their customers with their consumption and transaction data in a portable, preferably machine readable format. A similar initiative has been launched in France by Fing (Fondation Internet Nouvelle Génération), which provides a web-based platform MesInfos,[47] for consumers to access their financial, communication, health, insurance and energy data that are being held by businesses. Both the UK and French platforms are outgrowths of ProjectVRM,[48] a US initiative launched in 2006 that provides a model for Vendor Relationship Management by individual consumers. Finally, the right to data portability suggested by the EC in the current proposal for reform of their data protection legislation aims at stimulating innovation through more efficient and diversified use of personal data, by allowing users "to give their data to third parties offering different value-added services" (EDPS, 2014).

Data portability may involve significant costs to those that need, want, or must implement portability in their (existing) data-driven services. These include, costs both for developing and maintaining the mechanisms for enhanced data access, and for complying with relevant regulations (see Chapter 5). This may raise questions about who should bear the costs for developing and maintaining these mechanisms.[49]

2.4. Key findings and policy conclusions

A global data ecosystem is emerging in which, more than ever before, data and analytic services are traded and used across sectors and across national borders. For the ICT industry alone, this represents a USD 17 billion business opportunity that is growing at more than 40% on average every year since 2010. As a result, top ICT companies are strengthening their position through acquisitions of young start-ups specialised in big data technologies and services and/or through collaboration with potential competitors (co-opetition) in open source projects such as Hadoop. IBM was the most active acquirer of big data companies in 2012, followed by Oracle.

The top ICT firms contributing to the Hadoop ecosystem are to a large extent companies registered in the United States, with the exception of Yahoo Japan, NTT Data and Fujitsu (Japan), SAP (Germany), Persistent Systems (India) and Acer (Chinese Taipei). Most of the top ICT firms in the Hadoop ecosystem are Internet and software firms. Nevertheless, some hardware firms, in particular IT equipment firms, are heavily involved in big data-related technologies as well. Semiconductor firms, such as Intel and AMD, are the exceptions.

But the economic impact of the global data ecosystem goes far beyond the market prospects of the ICT industry, which mainly supplies goods and services for data collection, processing, and analysis. The data ecosystem involves a wide range of different types of actors with different business models and technologies. Besides ISPs and IT infrastructure providers, this includes in particular data service providers, analytic service providers, and data-driven entrepreneurs, many of these are start-ups. Many also act as users and producers of data and analytics, which suggests that the data ecosystem is a logical continuation of Web 2.0.

The global data ecosystem involves global value chains (GVCs), in which companies increasingly divide up their data related processes and locate productive activities in many countries. Figures on the distribution of data-driven services are not known. However, analysis of the world's top Internet sites suggests that data-driven services may be concentrated in the United States, which alone accounted for almost 60% of all top sites hosted in the OECD area in 2013, or more than 50% of all top sites hosted in OECD area plus Brazil, China, Colombia, Egypt, India, Indonesia, Russia, and South Africa taken together. The concentration of sites in the United States is most likely related to its well-functioning co-location and backhaul market, which reinforce the flourishing data ecosystem in that country. Statistics on trade in ICT-related services suggest further that the largest exporters of ICT service in 2013 – India, Ireland, the United States, Germany, the United Kingdom, and China – are more likely to be the largest destinations of cross-border data flows. As a consequence, the leading OECD importers of ICT-related services are also the major sources for trade-related data, and they include in particular the United States and Germany.

The characteristics of the global data ecosystem could create opportunities for BEPS through aggressive tax planning by multinational enterprises; this involves making use of gaps in the interaction of different tax systems to artificially reduce taxable income or shift profits to low-tax jurisdictions. What makes such action possible is the data ecosystem's ability to challenge the current paradigm used by tax authorities to determine where tax-relevant economic activities are carried out and value is created. Therein lies the difficulty in i) measuring the monetary value of data, ii) determining data ownership, and iii) acquiring a clear picture of the global distribution and interconnectedness of data-driven services..

The data ecosystem contains a rich mix of points of control that are distributed across all its layers, which however differ significantly across sectors. The exploitation of these points of control can raise serious competition and consumer protection concerns when they lead to the reduction of consumer choice, and anticompetitive behaviour. Lack of interoperability and vendor lock-in are two major risks through which points of controls can be exploited. In the area of cloud computing, the lack of open standards is still a huge problem, in particular in the area of platform as a service (PaaS). But points of control in the entrepreneurial layer also exist. These include, for example, data and walled gardens (i.e. closed proprietary platforms) based on multi-sided markets.

Analysis of points of control underlines the importance of (open) standards related to APIs and data formats. Lack of interoperability is among the most challenging barriers to the reuse of data and data-driven services in the data ecosystem. This is especially the case where data are not provided in a machine readable format and thus cannot be reused across IT systems. Individuals (consumers) also play an important role in promoting the free flow of their personal data across organisations if data portability is possible. Government and private sector initiatives promoting data portability are therefore

contributing to the free flow of data across organisations, and in so doing are strengthening the participation of individuals in DDI processes.

Characteristics of the global data ecosystem may also challenge the traditional approach employed by competition authorities to assess potential abuses and harms of market dominance and mergers. Challenges include: i) defining the relevant market, ii) assessing the degree of market concentration, and iii) ascertaining potential consumer detriment. Policy makers should encourage dialogue between competition, privacy and also consumer protection authorities, so that i) potential consumer harms due to DDI are taken into account, ii) synergies in enforcing rules controlling privacy violations, anticompetitive practices and mergers are unleashed, and iii) firms' incentives to compete with privacy-enhancing goods and services are increased.

Barriers to the open Internet, whether legitimate or not, can limit the effects of DDI. Some of these barriers may be technical, such as IP package filtering, or regulatory, such as "data localisation" requirements. They may result from business practices or government policies. Some of these have a legal basis, such as privacy and security (see Chapter 5), as well as the protection of trade secrets and copyright (OECD, 2015). However, these barriers can have an adverse impact on DDI – for example, if they limit trade and competition. Governments looking to promote DDI in their countries should consider further the OECD (2011b) *Council Recommendation on Principles for Internet Policy Making* as well as ongoing OECD work to develop better understanding of the characteristics and the social and economic impacts of an open Internet.

Annex – OECD (1985) Declaration on Transborder Data Flows

(Adopted by the Governments of OECD Member countries on 11th April 1985)

Rapid technological developments in the field of information, computers and communications are leading to significant structural changes in the economies of Member countries. Flows of computerised data and information are an important consequence of technological advances and are playing an increasing role in national economies. With the growing economic interdependence of Member countries, these flows acquire an international dimension, known as Transborder Data Flows. It is therefore appropriate for the OECD to pay attention to policy issues connected with these transborder data flows.

This declaration is intended to make clear the general spirit in which Member countries will address these issues.

In view of the above, the GOVERNMENTS OF OECD MEMBER COUNTRIES:

- Acknowledging that computerised data and information now circulate, by and large, freely on an international scale;
- Considering the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data and the significant progress that has been achieved in the area of privacy protection at national and international levels;
- Recognising the diversity of participants in transborder data flows, such as commercial and non-commercial organisations, individuals and governments, and recognising the wide variety of computerised data and information, traded or exchanged across national borders, such as data and information related to trading activities, intracorporate flows, computerised information services and scientific and technological exchanges;
- Recognising the growing importance of transborder data flows and the benefits that can be derived from transborder data flows; and recognising that the ability of Member countries to reap such benefits may vary;
- Recognising that investment and trade in this field cannot but benefit from transparency and stability of policies, regulations and practices;
- Recognising that national policies which affect transborder data flows reflect a range of social and economic goals, and that governments may adopt different means to achieve their policy goals;
- Aware of the social and economic benefits resulting from access to a variety of sources of information and of efficient and effective information services;
- Recognising that Member countries have a common interest in facilitating transborder data flows, and in reconciling different policy objectives in this field;

- Having due regard to their national laws, do hereby DECLARE THEIR INTENTION TO:

1. Promote access to data and information and related services, and avoid the creation of unjustified barriers to the international exchange of data and information;
2. Seek transparency in regulations and policies relating to information, computer and communications services affecting transborder data flows;
3. Develop common approaches for dealing with issues related to transborder data flows and, when appropriate, develop harmonised solutions;
4. Consider possible implications for other countries when dealing with issues related to transborder data flows.

Bearing in mind the intention expressed above, and taking into account the work being carried out in other international fora, the GOVERNMENTS OF OECD MEMBER COUNTRIES:

Agree that further work should be undertaken and that such work should concentrate at the outset on issues emerging from the following types of transborder data flows:

1. Flows of data accompanying international trade;
2. Marketed computer services and computerised information services; and
3. Intracorporate data flows.

The GOVERNMENTS OF OECD MEMBER COUNTRIES AGREED to co-operate and consult with each other in carrying out this important work, and in furthering the objectives of this Declaration.

Notes

1 This includes the market for technologies and services related to data storage, which is expected to be the fastest growing segment, followed by networking, and services.

2 This chapter is partly based on a follow-up study to TNO (2013), which was provided to the OECD as a contribution by the government of the Netherlands. To allow for an extensive investigation and detailed mapping of developments, TNO employed for the case studies a combination of top-down and bottom-up approaches. The case studies focus on a specific topic as a starting point of departure, and then the network exploration reaches beyond that initial domain in search of actors and markets that span the boundaries between sectors.

3 The notion of "entrepreneur" is to be understood here in a broader sense to include not only start-up entrepreneurs, but also civic entrepreneurs, who are engaged in social innovation, as well as public servants who are innovating in the public sector to give few examples. Ries (2011) discusses this broader notion of "entrepreneur" in more detail.

4 Data analytics is also used by ISPs for timely data transmission and for guaranteeing the delivering time of sensitive data even in crowded networks, through for example quality of service (QoS).

5 As described further below, many actors – including IT infrastructure providers, data providers, analytic service providers and data-driven entrepreneurs – are contributing to the development of open source software tools such as Hadoop and R, and are also generating, sharing or selling their data to third parties that can reuse the data for the development of new services.

6 The extent to which the data ecosystem could be referred to as the Web 4.0 (Web 3.0 being the Semantic Web) is left to the reader to decide.

7 Of the 100 randomly selected start-ups focusing on "big data" or "big data analytics" analysed by Hartmann et al. (2014), 70 businesses had a B2B business models, while 17 businesses built their businesses solely on a B2C model. The remaining 13 businesses used both models, B2B and B2C.

8 In January 2012 for example, Orange signed an agreement with Mediamobile, allowing it to use FMD data for its traffic information service V-Trafic – see www.traffictechnologytoday.com/news.php?NewsID=36182.

9 As Dumbill (2012c) explains: "Practical big data implementations don't in general fall neatly into either structured or unstructured data categories. You will invariably find Hadoop working as part of a system with a relational or MPP database."

10 "The MIT License is a permissive license that is short and to the point. It lets people do anything they want with your code as long as they provide attribution back to you and don't hold you liable. jQuery and Rails use the MIT License." (See http://choosealicense.com/).

11 The BSD License is "a permissive license that comes in two variants, the BSD 2-Clause and BSD 3-Clause. Both have very minute differences to the MIT license." (See http://choosealicense.com/licenses/).

12 "The Apache License is a permissive license similar to the MIT License, but also provides an express grant of patent rights from contributors to users. Apache, SVN, and NuGet use the Apache License." (See http://choosealicense.com/.)

13 "The GPL (V2 or V3) is a copyleft license that requires anyone who distributes your code or a derivative work to make the source available under the same terms. V3 is similar to V2, but further restricts use in hardware that forbids software alterations. Linux, Git, and WordPress use the GPL." (See http://choosealicense.com/.)

14 Clouds are sometimes also classified as private, public, or hybrid, according to their ownership and management control mechanisms.

15 These include e.g. Aristotle, LexisNexis, DocuSearch, Experian, Merlin Data, Pallorium. It is interesting to note that a combination of several data types – such as address, date of birth, social security number, credit record and military is estimated to cost around USD 55.

16 The public sector in the United States employed on average 1.6 database administrators per 1 000 employees in 2011.

17 On his first day in office, US President Obama announced his strategy for "open government", and the European Commission recently launched its Open Data Portal (Veenstra and en Broek, 2013).

18 These included Index Ventures and Khosla Ventures, SV Angel, Yuri Milner's Start Fund, Stanford Management Company, PayPal Founder Max Levchin; Google Chief Economist Hal Varian; and Applied Semantics' Co-Founder and Factual Chief Executive Officer Gil Elbaz.

19 The whole game is like an ongoing experiment. Foldit was successfully used to remodel the backbone of a computationally designed enzyme that catalyses the Diels-Alder reaction, which brings together two small molecules to form a particular kind of bond that the scientists were interested in making (see www.nature.com/nbt/journal/v30/n2/full/nbt.2109.html).

20 Walmart Labs is developing a number of (internal) solutions such as Social Genome, which allows Walmart to reach to potential customers, including friends of direct customers, who have mentioned specific products online, to provide discounts on these exact products. Social Genome builds on public data from the web (including social media data) as well as Walmart's proprietary data such as its customer purchasing and contact data. "This has resulted in a vast, constantly changing, up-to-date knowledge base with hundreds of millions of entities and relationships" (Big Data Startups, 2013).

21 For example, when population data from different sources are linked to health-sector data, some causes of illness can be better understood that could hardly be explained otherwise. An example is the analysis of environmental determinants of illnesses linked to nutrition, stress and mental health (OECD-NSF, 2011).

22 The 2012 Technology Foresight Forum (the Foresight Forum), held on 22 October 2012, highlighted the potential of big data analytics as a new source of growth. It put big data analytics in the context of key technological trends such as cloud computing,

smart ICT applications and the Internet of Things. It focused on the socioeconomic implications of harnessing data as a new source of growth and looked at specific areas: science and research (including public health), marketing (including competition) and public administration (see http://oe.cd/tff2012).

23 In a strategy referred to as "follow the moon", for example, companies such as Google are automatically and seamlessly shifting computing operations around the globe so computing heavy operations are done at night when the temperature is lower and cooling costs are cheaper (Higginbotham, 2009).

24 "For this analysis, the generic top-level domains were omitted from the list, as there is no reliable public data as to where the domains are registered. Out of the one million top sites, 948 00 were scanned, 474 000 were generic top-level domains, 40 000 had no identifiable host country, around 4 000 had no identifiable domain, just an IP-address. The remaining 429 000 domains were analysed and their hosting country identified. For each country the percentage of domains hosted in the country were [sic] identified" (OECD, 2014b; see also Pingdom, 2012).

25 This section is in part adapted from OECD, 2014d.

26 For example, while economic experiments and surveys in the United States indicate that individuals are willing to reveal their social security numbers for USD 240 on average, the same data sets can be obtained for less than USD 10 from data brokers in the United States such as Pallorium and LexisNexis.

27 The Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines), for example, recommends that individuals should have "the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him within a reasonable time; […] and d) to challenge data relating to him". These rights of the data subject are far-reaching and limit any possibility of exclusive right to the storage and use of personal data [by the data controller].

28 Mashups are web applications that use and combine content from different sources, including but not limited to web documents such as web pages and multimedia content; data such as cartographic and geographic data; and applications converter, communication, and visualisation tools.

29 However, the analysis undertaken here does not go much into the details of all relevant interaction, as does that undertaken by Clark (2012), which would have been necessary for a comprehensive control point analysis. Further studies following this method are therefore recommended for the future.

30 In many fields, competitive differentiation is most likely gained via proprietary data, data capturing interfaces, vertical analytics and visualisation. When speed is a differentiating asset, the data infrastructure will be essential as well. This is the case in high-frequency trading where hundreds of millions of dollars are being invested by companies in proprietary fibre optic cables to gain a milliseconds edge on their stock trading competitors.

31 This is in particular relevant when the collection of historical data series increases the value of the data. When these historical data are not portable, there is a (soft) lock-in, because the user loses this value. This is the case for instance in the agricultural

sector, where building a historical database provides increasing value and users are locked in and cannot move their data to another provider.

32 This section is in part adapted from OECD, 2014a.

33 OECD,2014a does not refer to "net neutrality", but takes the position that "the actions of market players can be described more accurately, with other terms, when considered across multiple countries that may not always use this term or do so with different interpretations".

34 Lock-in is still an attractive business strategy to maximise profit. As Swire and Lagos (2013) have argued, "the ability to attract users to a software service, and keep them there in at least some instances, is an important incentive for innovation and new entrants". Depending on their market and their market power, businesses could however develop a critical point of control though their lock-in capacity, that if exploited could raise consumer protection and competition issues.

35 Netflix, for example, uses Amazon's Web Services (AWS) for computing and storage (over 1 petabyte). Almost all of Netflix's information technology services run on AWS. Additionally, Netflix uses the services from Aspera to manage its data in Amazon's cloud. Netflix relies heavily on Amazon's infrastructure and, in the process, is one of Amazon's biggest customers. Simultaneously, Amazon is also a competitor in the on-demand video market with its Amazon Prime services, and Netflix is supporting the development of "an ecosystem that could lead to more competition for Amazon in the long term." Coincidentally, the adoption of these technologies by other cloud infrastructure providers would make it easier for Netflix to migrate to a provider other than Amazon (see King, 2013).

36 http://mashable.com/2010/09/17/google-voice-app-store-return/.

37 As Paul (2010) explains: "Many companies in technical fields attempt to collect as many broad patents as they can so that they will have ammunition with which to retaliate when they are faced with patent infringement lawsuits." For more on IP strategies (see OECD, 2015).

38 Markets featuring a series of disruptive innovations can lead to patterns in which firms rise to positions of temporary monopoly power but are then displaced by a competitor with superior innovation.

39 For example, HTML5 has tags for showing video on a web page or allowing users to drag and drop elements within the browser window.

40 Music purchases are commonly free of DRM restrictions and can played on nearly any device, regardless of platform. Downloaded video content, however, is almost always tied to one platform and cannot be viewed on others (OECD, 2013b).

41 See www.iclarified.com/entry/comments.php?enid=22914&laid=33#commentsanchor, accessed 15 May 2015.

42 This section benefited from the OECD Competition Committee hearings on the digital economy. Two hearings were held, in October 2011 and February 2012. OECD, 2013e includes an executive summary, an issues note by the Secretariat, a summary of each hearing, papers from panellists Eric Brousseau and Tim Wu and written submissions from: France, Japan, Norway, Poland, Turkey and Russia.

43 This observation has been confirmed in the OECD (2013f) work on competition in the digital economy undertaken under the first phase of the OECD (2013g) horizontal project on New Sources of Growth: Knowledge-based Capital (KBC 1). The conclusion reached was that markets characterised by the economic properties described above (increasing returns to and economies of scale and scope, paired with multi-sided markets and network effects) can lead to a "winner takes all" outcome where monopoly is the nearly inevitable outcome of market success.

44 It should be noted that in 2014 the UK Office of Fair Trading (OFT) became part of the Competition and Markets Authority (CMA), following reform of the competition system in the United Kingdom.

45 The study estimates the following effects: Brazil (-0.8%), the EU (-1.1%), India (-0.8%), Indonesia (-0.7%), Korea (-1.1%).

46 As an example, the Swedish standardisation committee "DIPAT" – SIS/TK 542, run by the Swedish Standards Institute (SIS), launched an initiative to work on national and European-level standardisation issues, linking and aligning the initiative with global efforts run by Subcommittee 38 of the Joint Technical Committee 1 of the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC JTC 1/SC 38). The goal is to assist in the development of harmonised, sustainable and well-designed standards.

47 See: http://fing.org/?-MesInfos-les-donnees-personnelles-&lang=fr.

48 See: http://cyber.law.harvard.edu/projectvrm/Main_Page.

49 The question is, should the data controller who will have to implement the mechanism pay, or the customers who request data portability, or the government that promotes the free flow of data across organisations and individuals?

References

Adner, R. (2006), "Match your innovation strategy to your innovation ecosystem", *Harvard Business Review*, April, http://pds12.egloos.com/pds/200811/07/31/R0604Fp2.pdf, accessed 12 June 2015.

Amazon (2009), "Amazon Elastic MapReduce Developer Guide API", 30 November, http://s3.amazonaws.com/awsdocs/ElasticMapReduce/latest/emr-dg.pdf, accessed 12 June 2015.

Angwin, J. (2010), "The web's new gold mine: Your secrets", *Wall Street Journal*, 30 July, http://online.wsj.com/article/SB10001424052748703940904575395073512989404.html.

Arthur D. Little, (2013), "Cloud from Telcos: Business distraction or a key to growth?", Arthur D. Little, www.adlittle.com/downloads/tx_adlreports/2013_TIME_Report_Cloud_from_Telcos.pdf, accessed 25 May 2015.

Assay, M. (2013), "VMWare: If Amazon wins, we all lose", *Readwrite,* http://readwrite.com/2013/03/01/vmware-if-amazon-wins-we-all-lose#awesm=~ohpgw6RGoncktJ, accessed 22 May 2015.

Bakhshi, H. and J. Mateos-Garcia (2012), "Rise of the Datavores: How UK businesses can benefit from their data", *Nesta,* 28 November, www.nesta.org.uk/publications/rise-datavores-how-uk-businesses-can-benefit-their-data.

Biesdorf, S., D. Court and P. Wilmott (2013), "Big data: What's your plan?", *McKinsey Quarterly,* McKinsey & Company, www.mckinsey.com/insights/business_technology/big_data_whats_your_plan, accessed 22 May 2015.

Big Data Startups (2013), "Walmart makes big data part of its DNA", http://smartdatacollective.com/bigdatastartups/111681/walmart-makes-big-data-part-its-social-media, accessed 22 May 2015.

Bonina, C. (2013), "New business models and the value of open data: Definitions, challenges, opportunities", RCUK Digital Economy Theme, www.nemode.ac.uk/wp-content/uploads/2013/11/Bonina-Opendata-Report-FINAL.pdf, accessed 12 June 2014.

Bruner, J. (2012), "Will data monopolies paralyze the Internet?", *Forbes*, 12 April, www.forbes.com/sites/jonbruner/2012/04/12/will-data-monopolies-paralyze-the-internet/.

Brynjolfsson, B. and A. McAffee (2012), "Big data: The management revolution", *Harvard Business Review*, October, http://hbr.org/product/big-data-themanagement-revolution/an/R1210C-PDF-ENG, accessed 12 June 2015.

Bunge, J. (2014), "Big data comes to the farm, sowing mistrust: Seed makers barrel into technology business", *The Wall Street Journal*, 25 February.

Capgemini Consulting (2013), "The open data economy: Unlocking the economic value by opening government and public data", www.capgemini.com/resources/the-open-data-economy-unlockingeconomic-value-by-opening-government-and-public-data, accessed 12 June 2014.

Chang, F. et al. (2006), "Bigtable: A distributed storage system for structured data", Google, appeared in Seventh Symposium on Operating System Design and Implementation (OSDI'06), November, http://research.google.com/archive/bigtable.html, accessed 24 May 2015.

Chen, L. et al. (2012), "Business intelligence and analytics: From big data to big impact", *MIS Quarterly*, Vol. 36, No. 4, pp. 1165-88.

Christensen, C.M. (1997), *The Innovator's Dilemma*, Harvard Business School Press, Boston.

Clark, D. (2012), "Control point analysis", 2012 TRPC Conference, 10 September, Social Science Research Network (SSRN), http://ssrn.com/abstract=2032124.

Cooper, J.C. (2013), "Privacy and antitrust: Underpants gnomes, the First Amendment, and subjectivity", *George Mason Law Review*, Rev. 1129 (2013), http://www.law.gmu.edu/assets/files/publications/working_papers/1339PrivacyandAntitrust.pdf, accessed 24 May 2015.

Criscuolo, P., N. Nicolaou and A. Salter (2012), "The elixir (or burden) of youth? Exploring differences in innovation between start-ups and established firms", *Research Policy*, Vol. 41, No. 2, pp. 319-333.

Cukier, K. (2010), "Data, data everywhere", *The Economist Special Report*, 25 February, www.economist.com/node/15557443.

Datameer (2013), "Hadoop Ecosystem: Who has the most connections", Datameer blog, http://datameer2.datameer.com/blog/wp-content/uploads/2013/01/hadoop_ecosystem_full2.png, accessed 12 June 2015.

Dean, J. and S. Ghemawat (2004), "MapReduce: Simplified data processing on large clusters", in Sixth Symposium on Operating System Design and Implementation (OSDI'04), December, San Francisco, http://research.google.com/archive/mapreduce.html, accessed 12 June 2015.

Duckett, C. (2014), "Computing experts call for repeal of copyrightable API decision", ZDnet, 10 November, www.zdnet.com/computing-experts-call-for-repeal-of-copyrightable-api-decision-7000035590/.

Dumbill, E. (2012a), "Microsoft's plan for big data", *O'Reilly Planning for Big Data*, http://oreilly.com/data/radarreports/planning-for-big-data.csp, accessed 12 June 2014.

Dumbill, E. (2012b), "Big data market survey", *O'Reilly Planning for Big Data*, http://oreilly.com/data/radarreports/planning-for-big-data.csp, accessed 12 June 2015.

Dumbill, E. (2012c), "Big data market survey: Hadoop solutions", *O'Reilly Strata*, http://strata.oreilly.com/2012/01/big-data-ecosystem.html, accessed 12 June 2015.

Dumbill, E. (2011), "Five data predictions for 2012", *O'Reilly Strata*, http://strata.oreilly.com/2011/12/5-big-data-predictions-2012.html, accessed 12 June 2014.

Dumbill, E. (2010), "The SMAQ stack for big data: Storage, MapReduce and Query are ushering in data-driven products and services", *O'Reilly Radar*, 22 September, http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html.

Dwyer, J. (2010). "Four nerds and a cry to arms against Facebook", *New York Times*, 11 May, www.nytimes.com/2010/05/12/nyregion/12about.html?_r=1.

ECIPE (European Centre for International Political Economy) (2014), "A friendly fire on economic recovery: A Methodology to estimate the costs of data regulations", *ECIPE Working Paper*, No. 02/2014, www.ecipe.org/media/publication_pdfs/WP22014.pdf, accessed 12 June 2015.

EDPS (2014), "Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy", European Data Protection Supervisor, March, https://secure.edps.europa.eu/EDPSWEB/edps/Home/Consultation/OpinionsC, accessed 24 May 2015.

EFF (2014), "Brief of a Amici Curiae Computer Scientists in Support of Petitioner", Electronic Frontier Foundation, 7 November.

EC (2013), "Study on business models for linked open government data", European Commission, http://ec.europa.eu/isa/documents/study-on-business-models-open-government_en.pdf, accessed 24 May 2015.

EC (2012), "Commission proposes a comprehensive reform of the data protection rules", Data Protection – Newsroom, European Commission, http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm, accessed 24 May 2015.

ESG (2012), "Boiling the ocean of control points in the Hadoop big data market", Enterprise Strategy Group, www.esg-global.com/blogs/boiling-the-ocean-of-control-points-in-the-hadoop-big-data-market/, accessed 24 May 2015.

Evans, D.S. (2011), "Antitrust Economics of Free", *Competition Policy International*, Spring 2011, http://ssrn.com/abstract=1813193.

Filistrucchi, L. et al. (2014), "Market definition in two-sided markets: Theory and practice", *Journal of Competition Law and Economics*, Vol. 10, No. 2, pp. 293-339.

Forbes (2013), "HTML5 vs. native mobile apps: Myths and misconceptions", 23 January, www.forbes.com/sites/ciocentral/2013/01/23/html5-vs-native-mobile-apps-myths-and-misconceptions/, accessed 24 May 2015.

FTC (2014), "Data brokers: A call for transparency and accountability", Federal Trade Commission, May, Washington, DC, https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf, accessed 24 May 2015.

FTC (2007), "Federal Trade Commission Closes Google/DoubleClick Investigation", 20 December, Federal Trade Commission, www.ftc.gov/news-events/press-

releases/2007/12/federal-trade-commission-closes-googledoubleclick-investigation, accessed 24 May 2015.

Gild (2014), "Intelligent sourcing", Gild, https://www.gild.com/score-candidates/, accessed 30 October 2014.

Ha, A. (2012), "DealAngel helps you find the best hotel deals — Not just the cheapest", *TechCrunch*, 17 April, http://techcrunch.com/2012/04/17/dealangel-helps-you-find-the-best-hotel-deals-not-just-the-cheapest/.

Harbour, P.J. and T. Koslov (2010), "Section 2 in a Web 2.0 world: An expanded vision of relevant product markets", *76 Antitrust J.L.*, pp. 769-94.

Harris, D. (2012), "Five low-profile start-ups that could change the face of big data", GigaOm, 28 January, http://gigaom.com/2012/01/28/5-low-profile-startups-that-could-change-the-face-of-big-data/.

Harris, D. (2011a), "As big data takes off, Hadoop wars begin", GigaOm, 25 March, http://gigaom.com/2011/03/25/as-big-data-takes-off-the-hadoop-wars-begin.

Harris, D. (2011b), "Why big data start-ups should take a narrow view", GigaOm, 28 March, http://gigaom.com/2011/03/28/why-big-data-startups-should-take-a-narrow-view/.

Hartmann, P. et al. (2014), "Big data for big business? A taxonomy of data-driven business models used by start-up firms", 27 March, Cambridge Service Alliance working paper, www.cambridgeservicealliance.org/uploads/downloadfiles/2014_March_Data%20Driven%20Business%20Models.pdf.

Heath, N. (2012), "Slow start for big data in Europe", *TechRepublic*, www.techrepublic.com/blog/european-technology/slow-start-for-big-data-in-europe/, accessed 24 May 2015.

Higginbotham, S. (2009), "Google gets shifty with its data center operations", GigaOm, 16 July, https://gigaom.com/2009/07/16/google-gets-shifty-with-its-data-center-operations/.

Kelly, J. (2013), "Hadoop pure-play business models explained", Wikibon, 17 December, http://wikibon.org/wiki/v/Hadoop_Pure-Play_Business_Models_Explained.

King, R. (2013), "Netflix brings Amazon web services closer", *The Wall Street Journal,* 28 July, http://blogs.wsj.com/cio/2013/07/28/netflix-brings-amazon-web-services-closer/.

Koehler, P., A. Anandasivam and M. Dan (2010), "Cloud services from a consumer perspective", *AMCIS 2010 Proceedings*, Paper 329, http://aisel.aisnet.org/amcis2010/329, accessed 24 May 2015.

Kommerskollegium (2014), "No transfer, no trade – The importance of cross-border data transfers for companies based in Sweden", January, www.kommers.se/Documents/dokumentarkiv/publikationer/2014/No_Transfer_No_Trade_webb.pdf, accessed 24 May 2015.

Koski, H. (2011), "Does marginal cost pricing of public sector information spur firm growth?", *ETLA Discussion Papers,* No. 1260, The Research Institute of the Finnish Economy, https://www.econstor.eu/dspace/bitstream/10419/87764/1/669255319.pdf, accessed 24 May 2015.

Lavalle, S. et al. (2010), "Analytics: The new path to value", *MIT Sloan Management Review*, p. 15, http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf, accessed 24 May 2015.

Lomas, N. (2013), "Handshake is a personal data marketplace where users get paid to sell their own data", *Techcrunch*, 2 September, http://techcrunch.com/2013/09/02/handshake/.

Loshin, D. (2002), "Knowledge integrity: Data ownership", Data Warehouse, 8 June, www.datawarehouse.com/article/?articleid=3052.

Metz, C. (2010), "Google blesses Hadoop with MapReduce patent license", *The Register*, 27 April, www.theregister.co.uk/2010/04/27/google_licenses_mapreduce_patent_to_hadoop/.

Microsoft (2011), "Microsoft expands data platform with SQL Server 2012: New investments for managing any data, any size, anywhere", *Microsoft News Center*, 12 October, www.microsoft.com/en-us/news/press/2011/oct11/10-12PASS1PR.aspx.

Miller, P. (2013), "Visualization, the key that unlocks data's value?", http://cloudofdata.com/2013/04/visualisation-the-key-that-unlocks-datas-value, accessed 24 May 2015.

Moore, F. (1993), "Predators and prey: A new ecology of competition", *Harvard Business Review*, May-June, http://blogs.law.harvard.edu/jim/files/2010/04/Predators-and-Prey.pdf, accessed 24 May 2015.

Muenchen, R. (2014), "The popularity of data analysis software", r4stats.com, http://r4stats.com/articles/popularity/, accessed 14 November 2014.

Musil, Steven (2011), "Report: Twitter buys TweetDeck for $40 million", CNET News, May, http://news.cnet.com/8301-1023_3-20065533-93.html, accessed 23 June 2015.

NAICS (2002), US North American Industry Classification System 2002.

NESSI (2012), "Big Data: A New World of Opportunities", Networked European Software and Services Initiative, www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf, accessed 25 May 2015.

NISO (2004), *Understanding Metadata*, NISO Press, National Information Standards Organization, www.niso.org/publications/press/UnderstandingMetadata.pdf, accessed 24 May 2015.

O'Brien, S.P. (2013), "Hadoop ecosystem as of January 2013 – Now an app!", Datameer, 15 January, www.datameer.com/blog/perspectives/hadoop-ecosystem-as-of-january-2013-now-an-app.html.

O'Dell, J. (2011), "In a world without tracking & cookies, can online commerce succeed?", mashable.com, 10 May, mashable.com/2011/05/10/buyosphere/.

OECD (2015), *Inquiries into Intellectual Property's Economic Impact?*, OECD Publishing, Paris, forthcoming.

OECD (2014a), "Connected televisions: Convergence and emerging business models", *OECD Digital Economy Papers*, No. 231, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jzb36wjqkvg-en.

OECD (2014b), "Cloud computing: The concept, impacts and the role of government policy", *OECD Digital Economy Papers*, No. 240, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxzf4lcc7f5-en.

OECD (2014c), "International cables, gateways, backhaul and international exchange points", *OECD Digital Economy Papers*, No. 232, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jz8m9jf3wkl-en.

OECD (2014d), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264221796-en.

OECD (2014e), *Addressing the Tax Challenges of the Digital Economy*, OECD/G20 Base Erosion and Profit Shifting Project, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264218789-en.

OECD (2013a), "Exploring the economics of personal data: A survey of methodologies for measuring monetary value", *OECD Digital Economy Papers*, No. 220, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k486qtxldmq-en.

OECD (2013b), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'big data'", in OECD, *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-12-en.

OECD (2013c), *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, OECD Publishing, Paris.

OECD (2013d), "The app economy", *OECD Digital Economy Papers*, No. 230, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3ttftlv95k-en.

OECD (2012a), *OECD Internet Economy Outlook 2012*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264086463-en.

OECD (2012b), "Market definition: Competition committee policy roundtable", DAF/COMP(2012)19, OECD Publishing, Paris, 11 October, www.oecd.org/daf/competition/Marketdefinition2012.pdf, accessed 24 May 2015.

OECD (2011a), "Internet intermediaries: Definitions, economic models and role in the value chain", in *The Role of Internet Intermediaries in Advancing Public Policy Objectives*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264115644-4-en.

OECD (2011b), *Recommendation of the Council on Principles for Internet Policy Making,* OECD Publishing, Paris, www.oecd.org/daf/competition/44445730.pdf, accessed 24 May 2015.

OECD (2009), "Two-sided markets: Competition committee policy roundtable", DAF/COMP(2009)20, OECD Publishing, Paris, 17 December, www.oecd.org/daf/competition/Marketdefinition2012.pdf.

OECD (2008), *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information,* OECD Publishing, Paris, www.oecd.org/internet/ieconomy/40826024.pdf, accessed 24 May 2015.

OECD-NSF, (2011) "OECD-NSF Workshop: Building a smarter health and wellness future", Summary of key messages, 15-16 February, internal working document.

OFT (2004), *Assessment of Market Power: Understanding Competition Law*, Competition Law Guideline, Office of Fair Trading, www.gov.uk/government/uploads/system/uploads/attachment_data/file/284422/oft402.pdf, accessed 24 May 2015.

Orrick (2012), *The Big Data Report*, Orrick, http://www.slideshare.net/CBInsights/big-data-report-31586014, accessed 25 May 2015.

Paul, R. (2010), "Google's MapReduce patent: What does it mean for Hadoop?", Arstechnica.com, 20 January, http://arstechnica.com/information-technology/2010/01/googles-mapreduce-patent-what-does-it-mean-for-hadoop/, accessed 24 May 2015.

Pingdom (2013), "The top 100 web hosting countries", 14 March, http://royal.pingdom.com/2013/03/14/web-hosting-countries-2013, accessed 20 April 2014.

Rao, L. (2013), "As software eats the world, non-tech corporations are eating start-ups", *TechCrunch*, http://techcrunch.com/2013/12/14/as-software-eats-the-world-non-tech-corporations-are-eating-startups/, accessed 24 May 2015.

Rao, L. (2011), "Index and Khosla lead $11m round in Kaggle, a platform for data modeling competitions", *TechCrunch*, 2 November, http://techcrunch.com/2011/11/02/index-and-khosla-lead-11m-round-in-kaggle-a-platform-for-data-modeling-competitions/, accessed 24 May 2015.

Redman, T. (2008), *Data Driven*, Harvard Business School Publishing, Boston, p. 25.

Reuters (2011), "*Financial Times* pulls its apps from Apple store", 31 August, www.reuters.com/article/2011/08/31/us-apple-ft-idUSTRE77U1O020110831, accessed 24 May 2015.

Ries, E. (2011), *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*, 13 September, First edition, Crown Business.

Russom, P. (2011) "Big data analytics", TDWI Best Practices Report, p. 24, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx, accessed 24 May 2015.

Schubert, L., K. Jefferey, and B. Neidecker-Lutz (2010), "The future of cloud computing: Opportunities for European cloud computing beyond 2010", public version 1.0, http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf, accessed 22 June 2014.

Shapiro, C. and H.R. Varian (1999), *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of government in a digital age", Computer and Communications Industry Association, October.

Strategy Analytics (2011), "One billion HTML5 phones to be sold worldwide in 2013", 7 December, www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=5145.

Suster, M. (2010), "Social networking: The future", TechCrunch, 5 December, http://techcrunch.com/2010/12/05/social-networking-future/.

Swire, P. and Y. Lagos (2013), "Why the right to data portability likely reduces consumer welfare: Antitrust and privacy critique", Maryland Law Review, Vol. 72.2, http://digitalcommons.law.umaryland.edu/mlr/vol72/iss2/1, accessed 24 May 2015.

Telefónica (2012), "Telefónica launches Telefónica dynamic insights – A new global big data business unit", Press release, Telefónica, 9 October, http://blog.digital.telefonica.com/?press-release=telefonica-launches-telefonica-dynamic-insights-a-new-global-big-data-business-unit.

The Economist (2014), "Digital disruption on the farm", 24 May, www.economist.com/news/business/21602757-managers-most-traditional-industries-distrust-promising-new-technology-digital.

The Economist (2012), "Know thyself", 15 December, www.economist.com/news/business/21568438-data-lockers-promise-help-people-profit-their-personal-information-know-thyself.

The Economist (2011), "Incentive prizes: Healthy competition", 10 April, www.economist.com/blogs/babbage/2011/04/incentive_prizes.

TNO (2013), "Thriving and surviving in a data-driven society", TNO, 24 September, http://publications.tno.nl/publication/34610048/xcv74S/TNO-2013-R11427.pdf.

Twitter (2012), "Changes coming in Version 1.1 of the Twitter API", 16 August, https://blog.twitter.com/2012/changes-coming-to-twitter-api, accessed 23 June 2015.

Ubaldi, B. (2013), "Open government data: Towards empirical analysis of open government data initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

Van der Berg, R. (2014), "The connected television debate in OECD countries", OECD Insights blog, OECD Publishing, Paris, 23 January, http://oecdinsights.org/2014/01/23/the-connected-television-debate-in-oecd-countries/.

van Veenstra, A.F. and T.A. van den Broek (2013), "Opening moves – Drivers, enablers and barriers in open data in a semi-public organization", in *M.A. Wimmer, M. Janssen and H.J. Scholl (eds.)*, EGOV 2013, LNCS 8074, pp. 50-61.

Watters, A. (2011a), "An iTunes model for data", O'Reilly Strata, http://strata.oreilly.com/2011/04/itunes-for-data.html. **[accessed when?]**

Watters, A. (2011b), "Scraping, cleaning and selling big data", O'Reilly Strata, http://strata.oreilly.com/2011/05/data-scraping-infochimps.html, accessed 24 May 2015.

Wireless Week (2011), "Native apps vs. HTML5-based apps: What does the future hold?", 7 September, www.wirelessweek.com/articles/2011/09/native-apps-vs-html5-based-apps-what-does-future-hold.

Woo, B. (2013), "A mind blowing big data experience: Notes from Strata", Forbes, 27 February, www.forbes.com/sites/bwoo/2013/02/27/a-mind-blowing-big-data-experience-notes-from-strata-2013/.

Yahoo News, (2013), "Twitter's new policies kill three more apps", 7 March.

Further reading

Brave, S. (2012), "We don't need more data scientists – Just make big data easier to use", GigaOm, https://gigaom.com/2012/12/22/we-dont-need-more-data-scientists-just-simpler-ways-to-use-big-data/, accessed 25 May 2015.

OECD (2013e), "The digital economy: Competition committee hearings", DAF/COMP(2012)22, OECD Publishing, Paris, 7 February, www.oecd.org/daf/competition/The-Digital-Economy-2012.pdf.

OECD (2013f), "Competition policy and knowledge-based capital", in OECD, *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-7-en.

OECD (2013g), *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193307-en.

OECD (2004), "518111 Internet Service Providers", Glossary of Terms, OECD Publishing, Paris.

Rochet J.-C. and J. Tirole (2006), "Two-sided markets: A progress report", *RAND Journal of Economics*, RAND Corporation, Vol. 37, No. 3, pp. 645-67, http://ideas.repec.org/a/bla/randje/v37y2006i3p645-667.html.

Taylor, R. (2012), "The Hadoop ecosystem, visualized in Datameer", Datameer blog, www.datameer.com/blog/uncategorized/the-hadoop-ecosystem-visualized-in-datameer.html, accessed 25 May 2015.

Chapter 3

How data now drive innovation

This chapter highlights the key drivers of data-driven innovation (DDI), today a widespread socio-economic phenomenon. It documents the key trends leading to the adoption of data and analytics across the economy, which are related to i) data generation and collection, ii) data processing and analysis, and iii) data-driven decision making. It also shows how the confluence of these trends is leading to the "industrialisation" of knowledge creation and a paradigm shift in decision making towards decision automation. The chapter then highlights the limitations of data-driven decision making, and concludes with a discussion of the key policy implications.

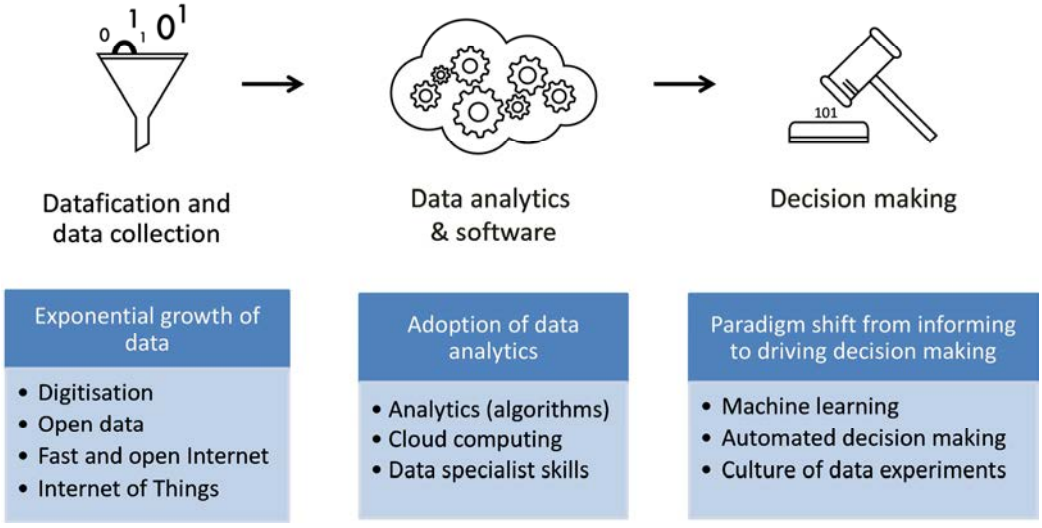
> *So how do we spot the future—and how might you? The seven rules that follow are not a bad place to start: […] 1. Look for cross-pollinators. […] 2. Surf the exponentials. […] 3. Favor the liberators. […] 4. Give points for audacity. […] 5. Bank on openness. […] 6. Demand deep design. […] 7. Spend time with time wasters. […]*. (Goetz, 2012)

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Data are an increasingly significant resource that can drive value creation and foster new industries, processes and products – data-driven innovation, or DDI. The importance of data however, both economically and socially, is not new. Many activities have long revolved round the analysis and use of data. Before the digital revolution, data were already used for scientific discovery and for monitoring business activities such as through accounting. There is also evidence that they were systematically collected and used in early history – for instance, as a means to keep information about the members of a given population (i.e. census).[1] In business, furthermore, concepts such as "business intelligence" (Luhn, 1958)[2] and "data warehousing" (Keen, 1978; Sol, 1987) emerged in the 1960s and became popular in the late 1980s when computers were increasingly used as decision support systems (DSSs). The financial sector is a popular example of the longstanding use of DSSs for (e.g.) detecting fraud and assessing credit risks (Inmon and Kelley, 1992).

That said, a confluence of three major socio-economic and technological trends is making DDI a new source of growth today: i) the exponential growth in data generated and collected, ii) the pervasive power of data analytics, and iii) the emergence of a paradigm shift in knowledge creation and decision making. These three trends are developing along the data value cycle introduced in Chapter 1 of this volume (Figure 1.7). Their confluence along the data value cycle has enabled the exploitation of data in ways never before possible. These three major trends are discussed further below, with the focus on key enabling factors, as illustrated in Figure 3.1.

Figure 3.1. **DDI: The data value cycle and confluence of key trends and enabling factors**



Note: Data specialist skills, key to adopting data analytics, are discussed in Chapter 6.

Understanding the key trends and enabling factors of DDI is crucial for governments to assess their economies' readiness to take advantage of this new source of growth. Economies in which these trends and factors are more prevalent are expected to be in a better position to benefit from DDI, although that does not mean that all factors need to be fully developed in order to realise the benefits. The global nature of the data ecosystem allows countries to benefit from DDI through data and analytics-related goods and

services produced elsewhere as discussed in Chapter 2 of this volume in more detail. However, it can be assumed that countries with enhanced capacities to both supply *and* use data and analytics will be in the best position to reap the fruits of DDI: A well-functioning supply side is a precondition for the development of a thriving data ecosystem, while a well-functioning demand side enables data-driven entrepreneurs to use data and analytics to innovate goods and services across the economy (see Chapter 2).

3.1. The exponential growth in data generated and collected

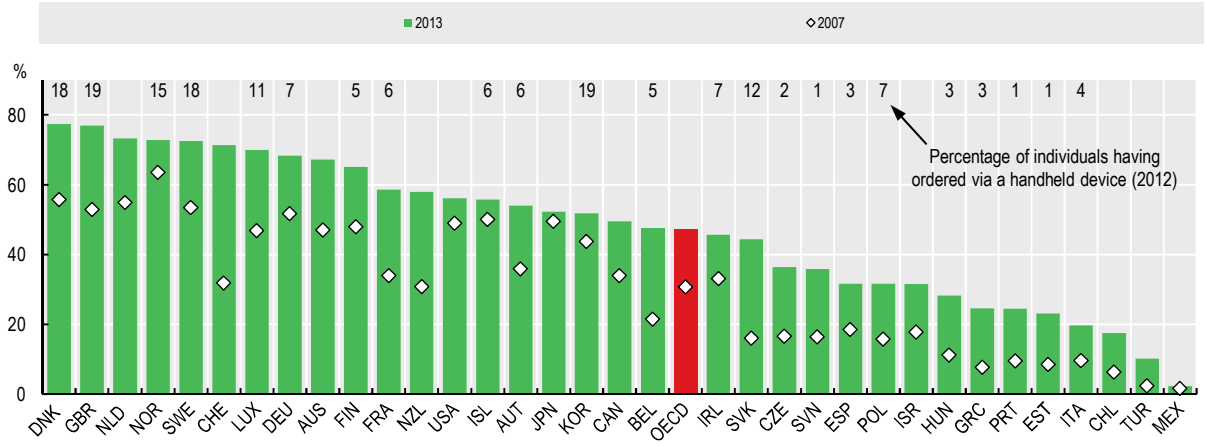
The first major trend driving DDI is the sheer growth in data volume. Measurement of the real total data generated, collected and stored is still speculative, but one source suggests that in 2010 alone, enterprises overall stored more than seven exabytes of new data on disk drives, while consumers stored more than six exabytes of new data (MGI, 2011). As a point of reference, one exabyte corresponds to one billion gigabytes and is equivalent, for example, to around 50 000 years of DVD-quality video. This growing storage has led to an estimated cumulative data volume of more than 1 000 exabytes in 2010, and some estimates suggest that that figure will multiply by a factor of 40 by the end of this decade, given the emerging Internet of Things (see section below) (IDC, 2012a).

The digitisation of nearly all media and the increasing migration of social and economic activities to the Internet (through Internet-based services such as social networks, e-commerce, e-health and e-government) have been two of the most important developments leading to the generation of unprecedented volumes of digital data across all sectors of the economy and in all areas of social life. In 2013, about half of the population in OECD countries, for example, had already purchased goods and services on line – thereby generating data that are increasingly used for personalised marketing, including product recommendation and personalisation (Figure 3.2).

In addition, the increasing deployment of connected devices through mobile and fixed networks captures an ever growing number of (offline) activities in the physical world. This process of transforming the world into processable and quantifiable data is sometimes referred to as "datafication", a portmanteau for "data" and "quantification" (Hey, 2004; Bertolucci, 2013; Mayer-Schönberger and Cukier, 2013).[3] The datafication of offline activities is resulting in an additional tidal wave of data. In 2013, there were almost 7 billion mobile subscriptions worldwide, of which roughly 15% were smartphones capable of running mobile device applications (apps), which collect and transmit a wide array of sensor data (Cisco, 2013; ITU, 2014). As the number of smartphones continues to grow, as well as the number of apps installed on these devices, more data can be expected to be generated, even if all apps are not actively used. In 2013, smartphone users on average installed around 30 apps on their smartphones, most of which are collecting data related to (for instance) communication, locations and personal accounts (OECD, 2014).

Figure 3.2. **The diffusion of online purchases, 2013 and 2007**

Percentage of individuals ordering goods or services online



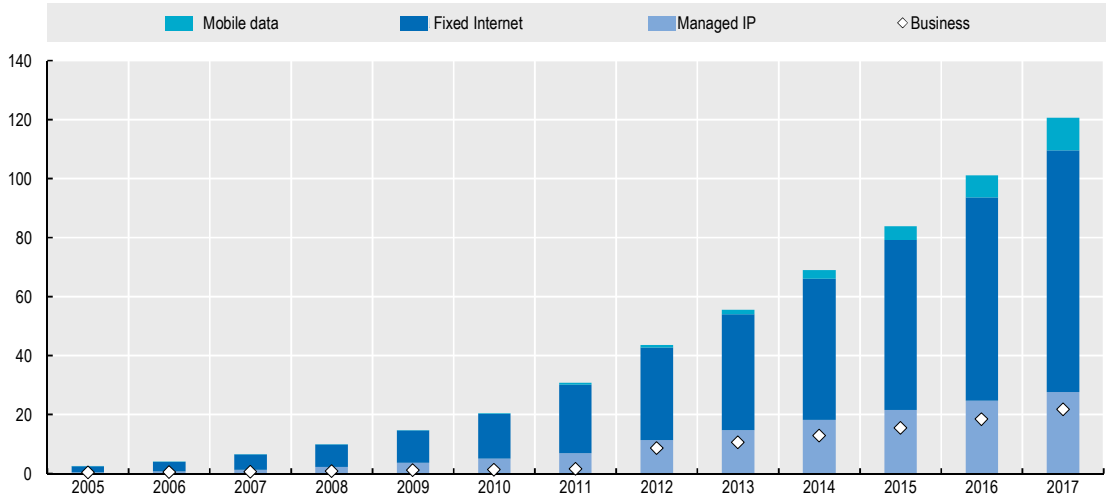
Note: For Australia, data refer to 2012/2013 (fiscal year ending in June 2013) instead of 2013. For 2007, data refer to 2006/2007 (fiscal year ending in June 2007), and to individuals aged 15 and over instead of 16-74 year-olds. For Canada, data refer to 2012 and relate to individuals who ordered goods or services over the Internet from any location (for personal or household use). For Chile, data refer to 2009 and 2012.For Japan, data refer to 2012 and to individuals aged 15-69 instead of 16-74 year-olds. For Israel, data refer to all individuals aged 20 and over who used the Internet for purchasing all types of goods or services. For Korea, the figure shows OECD estimates based on the Survey on the Internet Usage 2012. Data refer to the population aged 12 or more. In 2013, the share of individuals buying via handheld devices reached 35.5%. For New Zealand, data refer to 2006 and 2012 and relate to individuals who made a purchase through the Internet for personal use, which required an online payment. For Switzerland, data refer to 2005 instead of 2007. For the United States, data originate from May 2011 and September 2007 PEW Internet Surveys and cover individuals aged 18 or more.

Sources: OECD (2014), *Measuring the Digital Economy,* OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on OECD ICT Database; Eurostat, Information Society Statistics and national sources, May 2014.

Apps have transformed smartphones into multi-purpose mobile devices that in 2013 have generated more than 1.5 exabytes (billions of gigabytes) of data every month worldwide. However, the growth in mobile data is not only driven by the use of smartphones or tablets, which are estimated to account for only half of total mobile traffic. An even faster-growing volume of data about (offline) activities in the physical world is being generated by what is called the Internet of Things (IoT): interconnected objects enabled by sensors and machine-to-machine communication (M2M). Overall, Cisco (2013) estimates that the amount of data traffic generated by all mobile devices will almost double every year, reaching more than 11 exabytes by 2017 (Figure 3.3). This "datafication" process will reach its tipping point once the volume of (fixed and mobile) M2M bypasses that of human data communication, signalling a new phase of DDI that today is only in its infancy even in the most advanced economies.

Figure 3.3. **Monthly global Internet Protocol (IP) data traffic, 2005-17**

In exabytes (billions of gigabytes)



Source: OECD based on Cisco (2013), "Cisco Visual Networking Index: Forecast and methodology", 2012-17, www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html (June 2014).

The following sections look at the key enabling factors for the exponential growth of data and the IoT. It should be pointed out that although the main enablers have been identified, further studies are needed to fully understand the social and economic effects and the policy implications of the IoT, which go far beyond the scope of this section. The key enabling factors for the exponential growth of data include:

- access to a fast and open Internet – enabling the free flow of data
- sensors and sensor networks – enabling the ubiquitous datafication of the physical world
- machine-to-machine communication – empowering data exchange in the Internet of Things.

Access to a fast and open Internet – enabling the free flow of data

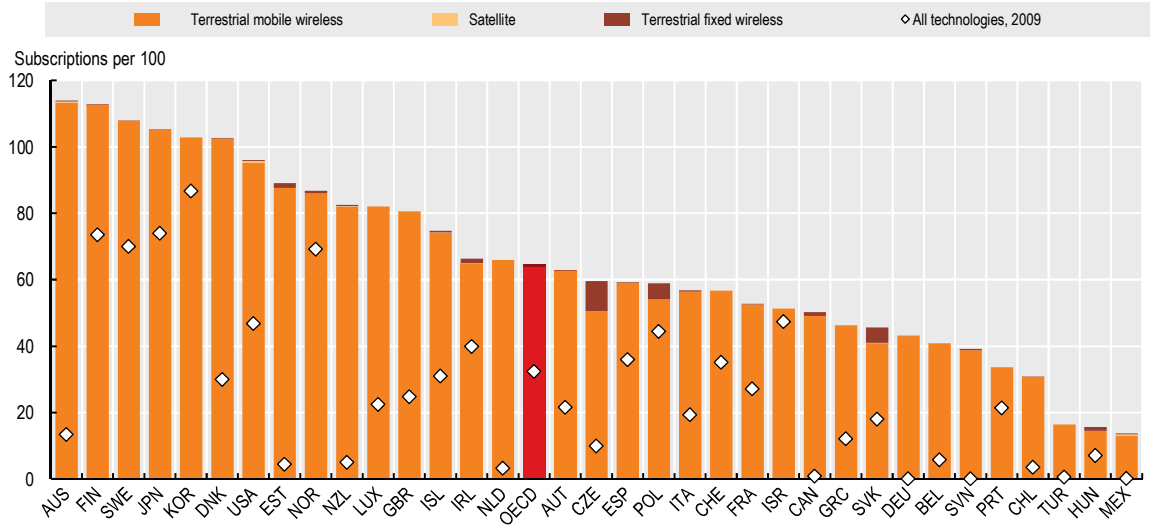
High-speed mobile broadband

The rapid diffusion of broadband is one of the most fundamental enablers of DDI. High-speed broadband is the underlying infrastructure for the exchange and free flow of data that are collected remotely through Internet applications and now increasingly through smart and interconnected devices. Where real-time applications are deployed, broadband networks enable timely data transmission (OECD, 2014a).[4] Mobile broadband in particular is essential, as mobile devices are becoming the leading means for data collection and dissemination. Moreover, high-speed mobile broadband is especially important to further improve connectivity in remote and less developed regions, where DDI could bring much needed (regional) growth and development (see Chapter 1). Within 10 years, between 2003 and 2013, fixed broadband penetration rates (subscribers per 100 inhabitants) in the OECD area have almost tripled, to reach around 30% of the OECD populations, but mobile broadband penetration rates have been more dynamic since surpassing fixed

broadband penetration rates in 2008. Since then, mobile broadband penetration rates have more than doubled, currently reaching around 70% in the OECD area.

The lowering of mobile access prices is the prime factor behind the explosion of mobile subscriptions (OECD, 2014b). In Australia, Finland, Sweden, Japan, Korea, and Denmark mobile penetration rates exceeded 100% in 2013 (Figure 3.4). Australia, which edged into first place after a 13% surge in smartphone subscriptions in the first half of 2013 – as well as Estonia, New Zealand, the Netherlands, the Czech Republic, and Canada – have experienced a boost in mobile subscriptions since 2009. Penetration is still at 40% or less in Portugal, Greece, Chile, Turkey, Hungary and Mexico; however, considering progress to date and the universal diffusion of standard mobile subscriptions, mobile broadband could well catch up in lagging economies as well (OECD, 2014b). For countries, broadband constitutes a *necessary*, *although not sufficient*, infrastructure related condition for DDI. Other factors, such as the (local) availability of data-driven services – and the related question of how well countries' co-location and backhaul markets function –, and an open Internet that enables non-discriminatory access and the free flow of data, are also essential to ensure that DDI takes root within national borders.

Figure 3.4. **OECD wireless broadband penetration, by technology, December 2009 and June 2013**



Note: Standard mobile broadband subscriptions may include dedicated mobile data subscriptions when breakdowns are not available. Israel: data for June 2010 instead of 2009.

Source: OECD (2014), *Measuring the Digital Economy,* OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on *OECD Broadband Portal*, www.oecd.org/sti/broadband/oecdbroadbandportal.htm, May 2014.

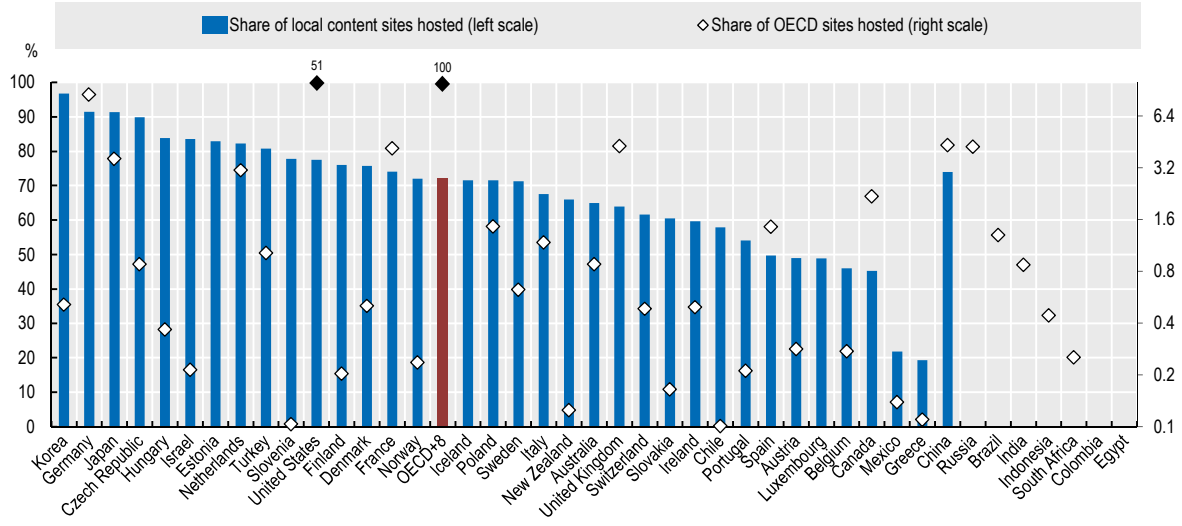
Co-location and backhaul markets

A recent OECD (2014c) study on "International Cables, Gateways, Backhaul and International Exchange Points" shows that the functioning of local markets for hosting and co-location has an effect on where digital local content (including data-driven services) is hosted. The study analyses the co-location of country code top-level domains (ccTLDs) such as ".fr" (France) and ".jp" (Japan) as identified in the Alexa One Million Domains (a list of the top million sites of the world).[5] The underlying assumption is that "if a larger portion of sites is hosted outside the country, it could indicate that the local market for hosting and co-location is not functioning efficiently" (OECD, 2014c).[6] The

analysis of local content sites hosted within countries shows that countries above the OECD average (e.g. Korea, Germany, Japan, Czech Republic, and Hungary) tend to conform to expectations that local content is hosted primarily within the country. Countries such as Greece, Mexico, Canada, Belgium, Luxembourg, Austria, Spain and Portugal have the lowest proportion of their most popular local content sites hosted domestically.

As the OECD (2014c) suggests, "it seems possible … that the market for co-location in Greece is unfavourable and content providers have not chosen a domestic location to host traffic. […] The factors at work in Greece are likely to be similar for Mexico, combined with the proximity to the United States, which has a well-functioning co-location and backhaul market". How well the co-location and backhaul market in the United States functions is indicated by the total number of sites hosted in the United States, which accounts for almost 60% of all top sites hosted in the OECD area in 2013, or more than 50% of all top sites hosted in the OECD area plus Brazil, People's Republic of China (hereafter 'China'), Colombia, Egypt, India, Indonesia, Russia and South Africa altogether (Figure 3.5). Grouping the European and Asian countries into regions may give a better perspective. In 2013, the United States accounted for 42% of all top sites hosted, while Europe hosted 31% of the world's top sites and Asia 11% (Pingdom, 2013). The number of top sites hosted strongly correlates with the number of co-location data centres (see Figure 1.5 in Chapter 1). This suggests that these top countries will be the main destinations for the global data flows on which DDI relies. Further analysis of the data reveals that for mid-income countries, the percentage of local content sites (and data centres) domestically hosted is correlated with the reliability of the electricity supply of that country (OECD, 2014c). This underlines "the importance of considering local energy supply when developing initiatives to enhance local backhaul and data centre markets" (OECD, 2014c).[7]

Figure 3.5. **Local content sites hosted in country, 2013**



Note: Based on the analysis of 429 000 ccTLD of the top one million sites. The remaining sites including the generic top-level domains were omitted from the list, as there is no reliable public data as to where the domains are registered.

Sources: Based on Pingdom, 2013; and www.datacentermap.com, accessed 27 May 2014.

The open Internet

There is still no widely agreed on definition of the open Internet; further studies are needed to develop a better understanding of its characteristics, and its social and economic impact. As highlighted in Chapter 2 of this volume, the openness of the Internet is a condition not only for information and knowledge exchange, but also for global competition among data-driven service providers. Most importantly, it is a vital condition for the nurturing of data-driven services that use and combine content (including data) from more than one source (i.e. mashups[8]) (Leipzig and Li, 2011). Many of these data-driven mashups draw on the activities of firms that have made some of their innovative services available via application programming interfaces (APIs), many of which are open and for free. As a result, a data ecosystem has emerged that is distributed around the world (see Chapter 2).

Ushahidi, Inc., based in Nairobi, Kenya, is an illustration. This non-profit software company relies on the open Internet, as it provides free and open source software and services based on available APIs from Internet firms such as Google and Twitter. One of its first products, created in the aftermath of Kenya's disputed 2007 presidential election, is used to collect eyewitness reports of violence via email and text messages; the locations are visualised on Google Maps. Since then, Ushahidi's data-driven services have been used during crises around the world; for example, in the aftermath of the 2010 earthquake in Haiti and the 2010 earthquake in Chile, it was used to locate the wounded.

As discussed in Chapter 2, barriers to the open Internet, whether legitimate or not, can limit the effects of DDI. Some of these barriers may be technical, such as IP package filtering, or regulatory, such as "data localisation" requirements, and they may be the results of business practices and government policies. Some of these have a legal basis such as privacy and security (see Chapter 5) as well as the protection of trade secrets and copyright (see OECD, 2015b). However, barriers erected through technologies, business practices and/or regulation can have an adverse impact on DDI – for example, if they limit trade and competition.

There is a common interest among countries to find a consensus on how to maintain a vibrant and open Internet and to exchange views on better practices. The OECD's High-Level Meeting on the Internet Economy on 28-29 June 2011 addressed the openness of the Internet and how best to ensure the continued growth and innovation of the Internet and the digital economy. The resulting draft communiqué, which led to the OECD (2011a) *Council Recommendation on Principles for Internet Policy Making*, contains a number of basic principles aimed to help ensure that the Internet remains open and dynamic, that it "allows people to give voice to their democratic aspirations, and that any policy-making associated with it […] promote[s] openness and [is] grounded in respect for human rights and the rule of law". The first five principles, listed below, are particularly relevant for the use of data. This is not to say that other principles are less important to DDI overall:

1. promote and protect the global free flow of information
2. promote the open, distributed and interconnected nature of the Internet
3. promote investment and competition in high-speed networks and services
4. promote and enable the cross-border delivery of services
5. encourage multi-stakeholder cooperation in policy development processes.

Sensors and sensor networks: Enabling the ubiquitous datafication of the physical world

The ubiquity of sensors is already reflected in the widespread use of smartphones, which account for roughly 15% of the 7 billion mobile subscriptions worldwide. This vast reach has its origins in technology's shift two decades ago, from electro-mechanical constructions to sensors and actuators built in silicon, in much the same way chips are built. That made possible the mass production of sensors, which today are embedded in far more than smartphones. Over 30 million interconnected sensors are estimated to be deployed worldwide today in areas such as security, health care, the environment, transport systems and energy control systems, and their numbers are growing by around 30% a year (MGI, 2011). Almost every adult in the OECD area today carries a number of sensors with them on a daily basis.[9]

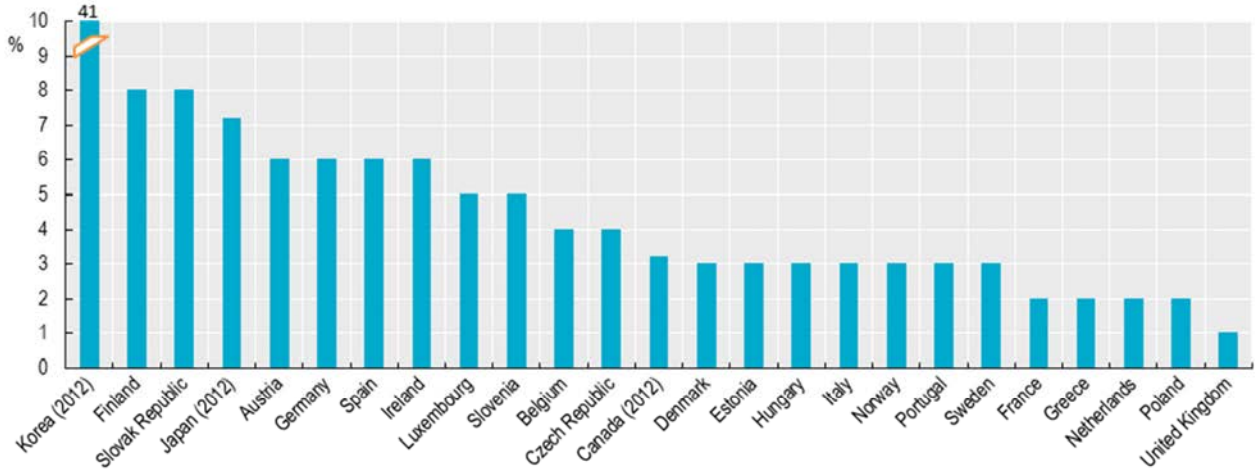
GPS sensors embedded in smartphones, for example, have enabled the generation of geo-locational data that are used in numerous apps and location-based services (mostly in real time), such as by online maps. In 2013, 68% of smartphone users in the OECD area have looked up directions or used a map on their smartphone, 18% more than in 2012; over 32% have searched for information about local businesses, and 14% (44% of those users) have actually visited the businesses afterwards (OECD, 2015a). Beyond its use in online maps and navigation systems, geo-locational data enable new services in areas such as shared mobility (see Chapter 9 of this volume) and multichannel retailing, to name but two.

Sensors have come down in price to such an extent that Apple's iPhone 5S now contains USD 3 worth of them, excluding camera (Hazard Owen, 2013). New sensors are being developed that can be used in novel ways, certain of which may seem unsettling to some. For example, Freescale, a semiconductor company, suggested a sensor that would be of use for gaming, fitness and health applications; at the same time however, it can measure emotions. This electrode-electrocardiogram and capacitive sensor can be integrated into a smart watch or fitness device. It can measure heart rate and sweat, and could thereby be used to measure physical activity. However, the heart rate and sweating are also involuntary indications of emotional states. The sensor would cost no more than USD 0.50 (Hazard Owen, 2013). More and more sensors are also becoming available for integration into different systems and devices. In industrial environments there are a great number of sensors for chemical and mechanical sensing, and more are currently being developed. This trend has been ongoing since the 1980s. A modern engine in a car cannot function without sensors; and it typically contains 50 sensors and sensor packages (Automotive Sensors Conference, 2015).

Looking across all sectors, however, sensor technologies remain underexploited. For instance, radio frequency identification (RFID) is still only adopted by a small set of businesses: less than 10% of all enterprises in OECD countries are using RFID technology. While in Korea 41% of enterprises with 10 or more employees have reported using RFID in 2012, the number is still only between 8% and 5% in Finland, the Slovak Republic, Japan, Austria, Germany, Spain and Ireland. In most other European countries the adoption rates are even lower (Figure 3.6).

Figure 3.6. **The diffusion of RFID in enterprises, 2011**

Percentages of enterprises employing 10 or more persons



Source: OECD (2014), *Measuring the Digital Economy,* OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361.

Some governments are increasingly deploying sensors as well, using them to measure everything from road conditions to trash collection. For example, the Netherlands government is deploying new fibre-based sensors to measure the stress dikes are undergoing, and integrating these into broadband fibre networks to neighbouring farms. Canada and Sweden use similar sensors to measure stress on the kilometre-long Île d'Orléans bridge near Quebec City and the Götaälvbron Bridge in Gothenburg. In the Swedish case it is a question of keeping the bridge safe and operational until a new bridge is built to replace it (Inaudi and del Grosso, 2008). Analysis of that data could boost smart transport and smart cities, which are discussed further in Chapter 9.

Machine-to-machine communication – Empowering data exchange in the Internet of Things

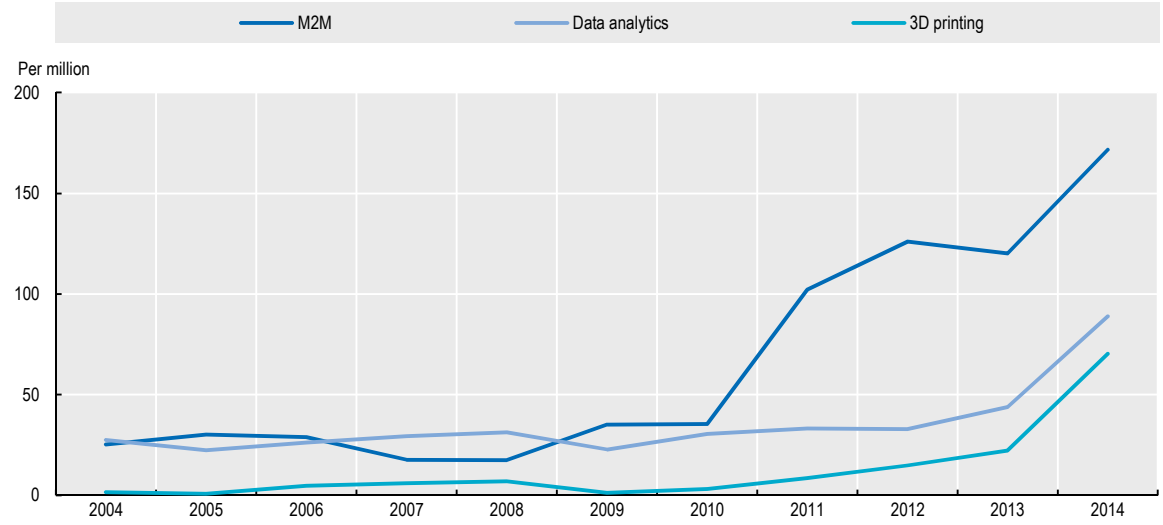
Today in OECD countries, an average family of four persons (including two teenagers) already has ten Internet-connected devices besides smartphones in and around the home, including tablets, printers and scanners, game consoles and increasingly smart TVs, smart meters, and Internet-connected cars. It is estimated that the average number of interconnected devices per household could reach 50 by 2022 (OECD, 2013a). In other words, the number of connected home devices in OECD countries would rise from over 1 billion today to 14 billion by 2022. This calculation only features OECD homes and does not take into account the growth in numbers of connected devices in industry, business, agriculture or public spaces, nor does it include devices in non-member economies. The McKinsey Global Institute (MGI, 2011) estimates that the number of connected smart devices will increase by more than 30% between 2010 and 2015, with the number of mobile-connected devices exceeding the world's population in 2012 (Cisco, 2013). Ericsson (2010) estimates that by 2025, as many as 50 billion devices will be online. That equates to 6 devices for each of the 8.1 billion people in the world by that time.

All these devices will exchange data to communicate with each other, a process known as machine-to-machine communication (M2M) (OECD, 2012b). Keyword text searches on international patent filings with the World Intellectual Property Organization

(WIPO) under the Patent Cooperation Treaty (PCT) provide evidence that M2M is rapidly increasing in importance when it comes to inventive activity (Figure 3.7). Following a sharp increase in 2011, more than 150 patent applications per million PCT patent applications were related to M2M in 2014, compared to 20 PCT patent applications in 2008.

Figure 3.7. **Patents on M2M, data analytics and 3D printing technologies, 2004-14**

Per million PCT patent applications including selected text strings in abstracts or claims



Note: Patent abstracts and/or claims were searched for the following: (a) M2M: "machine to machine" or "M2M"; (b) Data analytics: "data mining" or "big data" or "data analytics"; (c) 3D printing: "3D printer" or "3D printing". 2014 is limited to data available before 31 May.

Source: OECD (2014), *Measuring the Digital Economy*, OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on the OECD PATSTAT database.

Innovative products based on M2M include, for example, smart meters that collect and transmit real-time data on energy (OECD, 2012a), and Internet-connected automobiles that are now able to transmit real-time data on the state of the car's components and environment (OECD, 2012b) (Both applications are discussed in more detail in Chapter 9.) Many of these connected devices are based on sensor and actuator networks that sense and exchange data through wireless links "enabling interaction between people or computers and the surrounding environment" (Verdone et al., 2008, cited in OECD, 2009).[10] These sensors can be regarded as "the interface between the physical world and the world of electrical devices, such as computers" as they measure multiple physical properties. Examples include electronic sensors, biosensors, and chemical sensors (Wilson, 2008). The counterpart is represented by actuators that function the other way round, i.e. whose tasks consist in converting the electrical signal into a physical phenomenon (e.g. displays for quantities measures by sensors such as speedometers, temperature reading for thermostats, but also those that control the motion of a machine).

The use of sensor technology is not what is new here. What is changing is that the data are now not only used in the machine but also shared more widely and combined with other data. In early sensor systems, such as in vehicle engines, the data were measured, processed, acted upon and discarded. More recently however, more and more

of the data generated are communicated and stored for further analysis. General Electric is one of the more visible companies promoting this development as an integral part of their vision of the Industrial Internet. Other industry initiatives such as those by Siemens on "networked manufacturing" highlight similar trends (*The Economist*, 2014). These initiatives see a future where machines are built with many different sensors that continuously collect and send data, which are then analysed and acted upon at a system-wide level.

The type of communication used can vary between wired and wireless, short or long range, low or high power, and low or high bandwidth (OECD, 2013a). A way to order the applications and technologies is to look at the geographic distribution and mobility that has to be supported by the M2M networks (Figure 3.8). An increase in mobility and dispersion comes at a cost to energy and bandwidth, meaning that the applications will likely need a bigger battery and can send fewer data than those devices that stay in one location.

Figure 3.8. **Machine-to-machine applications and technologies, by dispersion and mobility**

	Geographically fixed	Geographically mobile
Geographically dispersed	*Application* – smart grid, smart metre, city, remote monitoring *Technology required:* PSTN, broadband, 2G/3G/4G, power line communication	*Application* – car automation, e-health, logistics, portable consumer electronics *Technology required* – 2G/3G/4G, satellite
Geographically concentrated	*Application* – smart home, factory automation, e-health *Technology required* – wireless personal area (WPA), networks, wired networks, indoor electrical wiring, Wi-Fi	*Application* – on-site logistics *Technology required* – Wi-Fi, WPAN

Source: OECD (2012), "Machine-to-Machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

Given the enormous amount of devices that will come on line in the coming years (50 billion devices by 2025), one important question is whether networks will be able to support all these devices. Network interactions initiated by humans are of a more intermittent character, with pauses between interactions. However, when people interact over networks, they expect interactions within less than 0.2 seconds, which limits the amount of data that can be sent to the user. On a 100 Mbit per second connection, this effectively reduces the amount of data that can be exchanged to 1.25 megabytes or less. M2M is different from traditional applications in that it is more upload focused and less "bursty". A smart meter may send many times more measurement data than it receives in control data over its lifetime, and it does so in a continuous stream. The same holds true with any other type of sensor. The data rates achieved very much depend on the data that are collected and the sampling rate.

Actual data rates also depend on the type of processing done on these data. In the case of an automobile, the data may be processed on board and as a result reduced in size to facilitate easier uploads of whatever data are relevant. As a consequence, however, an automobile would need adequate on-board processing power and sufficient energy supply. In other application cases, the data must first be uploaded because there is no local processing power available, or they need to be combined with other data before they becomes useful. The time sensitivity of data is another concern. If there are real-time feedback loops (e.g. smart meters), the data should be sent uncompressed; absent such loops, "lossless compression" can save bandwidth. In the case of real-time data, the data will need to be streamed.

Furthermore, the increasing deployment of interconnected devices will require governments to address the issue of migration to a new system of Internet addresses (IPv6). The current IPv4 addresses are essentially exhausted, and mechanisms for connecting the next billion devices are urgently needed. IPv6 is a relatively new addressing system that offers the possibility of almost unlimited address space, but adoption has been slow. M2M also raises regulatory challenges related to opening the access to mobile wholesale markets to firms not providing public telecommunication services; there are also numbering and frequency policy issues (see Box 3.1).

Box 3.1. **M2M and regulatory barriers to data-driven mobile applications**

Machine-to-machine communication (M2M) is an enabler of DDI in many industrial applications and services, including logistics, manufacturing, and even health care. However, a major barrier for the M2M-enabled mobile applications (and users) is the lack of competition once a mobile network provider has been chosen. The problem is the SIM (subscriber identity module) card, which links the device to a mobile operator. By design, only the mobile network that owns the SIM card can designate which networks the device can use. In mobile phones the SIM card can be removed by hand and changed for that of another network. But when used in cars or other machines it is often soldered, to prevent fraud and damage from vibrations. Even if it is not soldered, changing the SIM at a garage, a customer's home, or on-site, costs between USD 100 and USD 1 000 per device.

Consequently, once a device has a SIM card from a mobile network, the company that developed the device cannot leave the mobile network for the lifetime of the device. Therefore, the million-device user can effectively be locked into 10- to 30-year contracts. It also means that when a car or e-health device crosses a border, the large-scale user is charged the operator's costly roaming rates. The million-device user cannot negotiate these contracts. It also cannot distinguish itself from other customers of the network (normal consumers) and is covered by the same roaming contracts.

There are many technological and business model innovations that a large-scale M2M user wants to introduce. However, at present it cannot do so in most countries because it would need the approval of its mobile network operator. Many innovations would bypass the mobile operator and therefore are resisted. The solution would be for governments to allow large-scale M2M users to control their own devices by owning their own SIM cards, something that is implicitly prohibited in many countries. It would make a car manufacturer the equivalent of a mobile operator from the perspective of the network. Removing regulatory barriers to entry in this mobile market would allow the million-device customer not only to become independent of the mobile network but also to create competition. This would yield billions in savings on mobile connectivity and revenue from new services.

Source: OECD (2012), "Machine-to-Machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

3.2. The pervasive power of data analytics

Data analytics is the second major development favouring DDI. The large volume of data generated by the Internet, including the IoT, has no value if no information can be extracted from the data. Data analytics refers to a set of techniques and tools that are used to extract information from data. These techniques and tools extract information from data by revealing the context in which the data is embedded and its organisation and structure. They help reveal the "signal from the noise" and with that, the data's "manifold hidden relations (patterns), e.g. correlations among facts, interactions among entities, relations among concepts" (Merelli and Rasetti, 2013; see also Cleveland, 1982 and Zins,

2007). There are a number of terms that are used (as synonyms) to refer to data analytics, some of which may include aspects that go beyond traditional data analysis.

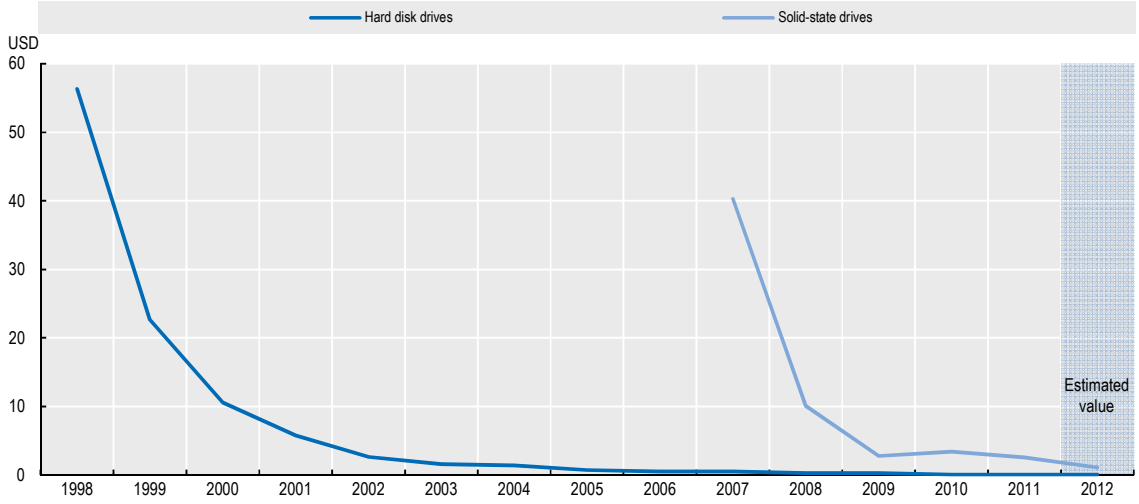
- *Data mining* refers to a set of techniques used to extract information patterns from data sets. It is often said to go beyond data analytics as it combines data analytic methods such as statistics and machine learning with data management technologies (e.g. SQL [structured query language] databases, distributed data management with tools such as Hadoop), and data pre-processing methods (data cleaning). The key aspect here, however, is the discovery of information patterns. Data mining is thus often used as a synonym for another term used more frequently in the past: *knowledge discovery*.
- *Profiling* refers to the use of data analytics for the construction of profiles and the classification of entities in specific profiles, both based on the attributes of these entities. The term is often used in cases where the profiled entities are individuals from which personal identifying information (PII) have been collected; credit scoring, price discrimination and targeted advertisement are typical examples of activities involving profiling. But the term can also be used where non-personal-related entities are being profiled (e.g. malware activities).
- *Business intelligence (BI)* was a term coined by Luhn (1958), who defines it by combining two Webster Dictionary definitions: that of i) intelligence: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal", and ii) business: "a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera." Today, BI refers to tools and techniques used to process data that have been previously stored in a database or data warehouse. The objective in BI is the creation of standard reports on a periodic basis, or the display of real-time business-related information on "management dashboards" highlighting key operation metrics for business management. BI typically focuses on databases for business reporting and monitoring. However, the boundaries between BI and data mining are blurring, as BI software vendors are increasingly offering products and services covering BI as well as data mining.
- *Machine or statistical learning* is a subfield in computer science, and more specifically in artificial intelligence. It is concerned with the design, development and use of algorithms that allow computers to "learn" – that is, to perform certain tasks while improving performance with every empirical data set it analyses. Machine learning involves activities such as pattern classification, cluster analysis, and regression (Mitchell et al., 1986, Duda, Hart and Stork, 2000; Russel and Norvig, 2009; Hastie, Tibshirani and Friedman, 2011; James et al., 2013).
- *Visual analytics* refers to techniques and tools used for data visualisation. They are used for gaining insights at a glance, including through interactive data exploration, and for communicating these insights to others (Unwin, Theus and Hofmann, 2006; Janert, 2010).

Data analytics is now becoming more affordable for start-ups and small and medium-sized enterprises (SMEs), and their adoption will intensify as the volume of data continues to grow. The growing interest in data analytics is also reflected in the number of scientific articles related to the topic. Within the past ten years (2004-2014) that number has, on average, grown by 9% each year (see Figure 1.9 in Chapter 1).

The adoption of data analytics has been greatly facilitated by the declining cost of data storage and processing. In the past, collection, storage and processing were expensive and for the most part available only to large corporations, governments and universities. With ICTs becoming increasingly powerful, ubiquitous and inexpensive, the exploitation of data is now becoming accessible to a wider population (OECD, 2013a). For example, storage costs in the past discouraged keeping data that were no longer, or unlikely to be, needed (OECD, 2011b). But storage costs have decreased to the point where data can be kept for long periods of time if not indefinitely. This is illustrated by the average cost per gigabyte of consumer hard disk drives (HDDs), which dropped from USD 56 in 1998 to USD 0.05 in 2012 – an average decline of almost 40% a year (Figure 3.9). With new generation storage technologies such as solid-state drives (SSDs), the decline in costs per gigabyte is even more rapid.

Figure 3.9. **Average data storage cost for consumers, 1998-2012**

In USD per gigabyte



Note: Data for 1998-2011 are based on average prices of consumer-oriented drives (171 HDDs and 101 SSDs) from M. Komorowski (www.mkomo.com/cost-per-gigabyte), AnandTech (www.anandtech.com/tag/storage) and Tom's Hardware (www.tomshardware.com). The price estimate for SSD in 2012 is based on DeCarlo (2011) referring to Gartner.

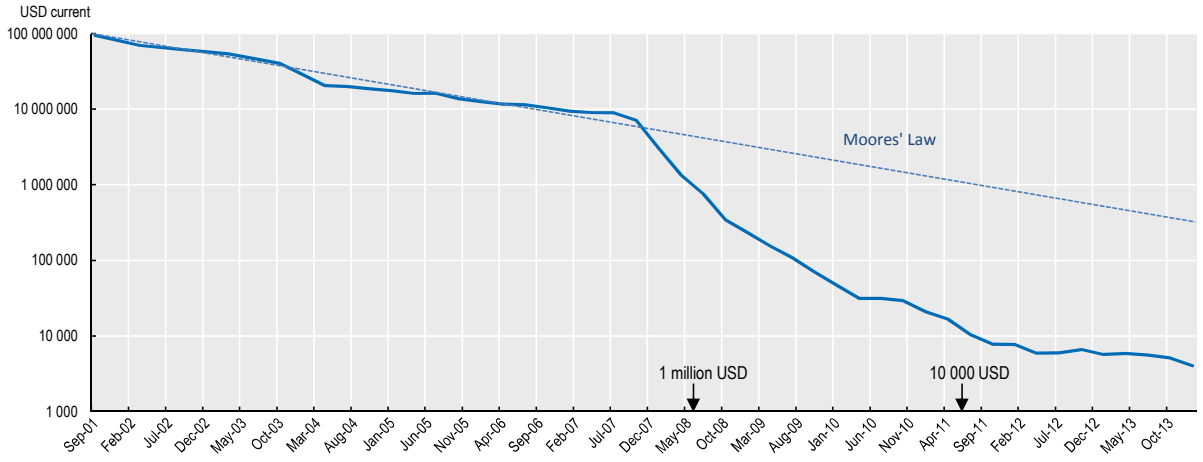
Source: Based on Royal Pingdom blog, December 2011.

The decline in data storage and processing costs is very much a reflection of Moore's Law, which holds that processing power doubles about every 18 months, relative to cost or size. However, as the evolution of the cost of DNA gene sequencing shows, other trends besides Moore's Law have largely contributed to the decreasing cost. The sequencing cost per genome has dropped at higher rates than Moore's Law would predict, from USD 100 million in 2001 to less than USD 6 000 in 2013 (Figure 3.10). Among the factors that have led to the dramatic cost reduction in data storage and processing, the following ones discussed further below should be highlighted:

1. improvements in algorithms and heuristic methods
2. the availability of open source software (OSS), covering the full range of solutions needed for data collection, storage, processing and analytics
3. the availability of computing power at massive scale thanks to cloud computing.

Figure 3.10. **Cost of genome sequencing, 2001-14**

Cost per genome in USD, logarithmic scale



Source: OECD (2014), *Measuring the Digital Economy*, OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on the National Human Genome Research Institute (NHGRI) Genome Sequencing Program (GSP) www.genome.gov/sequencingcosts/.

Algorithms, heuristic methods and data processing techniques

Significant progress has been made in the development of algorithms and heuristic methods to process and analyse large data sets. It comes as no surprise that Internet firms, in particular providers of web search engines, have been at the forefront of the development and use of techniques and technologies for processing and analysing large volumes of data. They were among the first to confront the problem of handling big streams of mainly unstructured data stored on the web in their daily business operation. Google in particular inspired the development of a series of technologies after it presented MapReduce, a programming framework for processing large data sets in a distributed fashion, and BigTable, a distributed storage system for structured data, in a paper by Dean and Ghemawat (2004) and Chang et al. (2006) respectively. Examples include Hadoop and CouchDB – both open source solutions (under the Apache License) – which have become the engine behind many of today's big data processing platforms (see Chapter 2, Box 2.3).

Some of the progress in data analytics is captured by patents. This includes, for example, the software method patent for a "system and method for efficient large-scale data processing" (US 7650331 B1) that covers the principle of *MapReduce* and that was awarded to Google by the United States Patent and Trademark Office (USPTO) in 2010. Looking at patent applications overall (Figure 3.7), one can observe the growth in the number of patent applications related to data analytics, in particular for "machine learning, data mining or biostatistics, e.g. pattern finding, knowledge discovery, rule extraction, correlation, clustering or classification" (IPC G06F 19/24). However, it is important to highlight that the numbers of data analytics patents can be misleading, for several reasons. Most importantly, numbers of patent applications and patents in data processing in general do not fully reflect ongoing innovation and therefore should be interpreted with caution, in particular when undertaking cross-country comparisons. This is because innovation in data processing is to a large extent embodied in software, for which the application and granting of a patent may vary significantly between countries.

Furthermore, much of the innovation in this field involves open source software (OSS), which is provided with free software licences such as the MIT License,[11] the BSD License,[12] the Apache License[13] and the GNU general public license (GPL v2 or v3).[14] While some of these free software licences provide an express granting of patent rights from contributors to users (e.g. Apache), others may include some form of patent "retaliation" clauses, which stipulate that some rights granted by the licence (e.g. redistribution) may be terminated if patents relating to the licensed software are enforced (e.g. the Apple Public Source License).[15]

The use of patents and copyright has raised a number of concerns in the data analytic community. For example, some have expressed concerns that the patent US 7650331 B1 on MapReduce awarded to Google could put at risk companies that rely entirely on the open source implementations of MapReduce, such as Hadoop and CouchDB (Paul, 2010; Metz, 2010a; 2010b). While such a concern may be justified given that Hadoop is widely used today, including by large companies such as IBM, Oracle and others as well as by Google, expectations are that Google "obtained the patent for 'defensive' purposes" (Paul, 2010).[16] By granting a licence to Apache Hadoop under the Apache Contributor License Agreement (CLA), Google has officially eased fears of legal action against the Hadoop and CouchDB projects (Metz, 2010b). In the area of copyrights, issues are related more to copyright protected data sources, which under some conditions may restrict the effective use of data analytics (Box 3.2).

Box 3.2. **Copyrights and data analytics**

Data analytics is leading to an "automation" of knowledge creation, with text mining constituting a key enabling technology (Lok, 2010). Based on early work by Swanson (1986), scientists are now further exploring the use of data analytics for automated hypothesis generation, and some have proposed analytical frameworks for standardising this scientific approach. Abedi et al. (2012), for example, have developed a hypothesis generation framework (HGF) to identify "crisp semantic associations" among entities of interest. Conceptual biology, as another example, has emerged as a complement to empirical biology, and is characterised by the use of text mining for hypothesis discovery and testing. This involves "partially automated methods for finding evidence in the literature to support hypothetical relationships" (Bekhuis, 2006). Thanks to these types of methods, insights are possible that otherwise would have been difficult to discover. One example is the discovery of adverse effects of drugs (Gurulingappa et al., 2013; Davis et al. 2013).

The potential for productivity gains in the creation of scientific knowledge are thus huge. However, questions have emerged about whether current copyright regimes are appropriately calibrated with regard to "automatic" scientific knowledge creation. According to the JISC (2012) analysis of the value and benefits of text mining, "the barriers limiting uptake of text mining appeared sufficiently significant to restrict seriously current and future text mining in UKFHE [UK further and higher education], irrespective of the degree of potential economic and innovation gains for society." Copyright has been identified as one these barriers, which has led to debates between the scientific community and the publishers of scientific journals (see OECD, 2015b).

Open source analytics

OSS applications that cover the full range of solutions needed for data processing and analysis (including visualisation) have contributed significantly to making data analytics accessible to a wider population. Many data processing and analytic tools that are now spreading across the economy as enablers of new data-driven goods and services were initially developed by Internet firms. Hadoop, the open source implementation of

Google's MapReduce, was already mentioned above. Another well-known example is R, a GPL-licensed open source environment for statistical analysis, which is increasingly used as an alternative to commercial packages such as SPSS and SAS. Today R is also an important part of the product portfolio of many traditional providers of commercial database and enterprise servers such as IBM,[17] Oracle,[18] Microsoft[19] and SAP,[20] which have started integrating R together with Hadoop into their product lines.

Measured by scholarly publications in Google Scholar, Muenchen (2014) estimates the popularity of statistical software including R to have grown significantly over past ten years, the assumption being that "the more popular a software package is, the more likely it will appear in scholarly publications as a topic and as a method of analysis". Muenchen's analysis of the number of articles for the most popular six statistics software from 1995 through 2012 suggests that the most popular statistics software (SPSS, SAS) is declining in popularity, while R is becoming more and more popular.[21] A survey undertaken by the data mining website KDnuggets (2013) confirms the trend that a large number of data analysts are using open source or free software for data analysis.[22]

Cloud computing: Providing super computing power as a utility

Cloud computing has played a significant role in increasing the capacity to store and analyse data. It has been described as "a service model for computing services based on a set of computing resources that can be accessed in a flexible, elastic, on-demand way with low management effort" (OECD, 2014d). Super computing power and data analytics are complementary resources needed to make sense of "big data", as analysis of large volumes of data requires huge computational resources – especially if the analysis needs to be performed in real time.

Cloud computing can be classified into three different service models according to the resources it provides: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS):[23]

- IaaS provides users with managed and scalable raw resources such as storage and computing resources
- PaaS provides computational resources (full software stack) via a platform on which applications and services can be developed and hosted
- SaaS offers applications running on a cloud infrastructure.

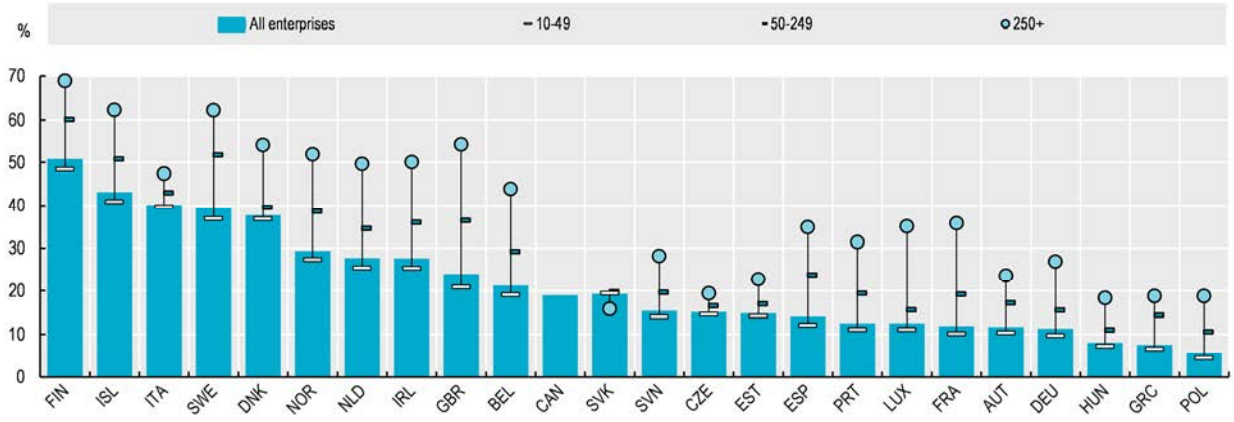
The benefits of cloud computing services can be summarised as efficiency, flexibility, and innovation. Cloud computing reduces computing costs through demand aggregation, system consolidation, and improved asset utilisation. In addition, it provides near-instantaneous increases and reductions in capacity in a pay-as-you-go model, which enables service users to act more responsively to customers' needs and demand without much initial investment in IT infrastructure (Kundra, 2011). All these factors lower the entry barriers of cloud-using markets for start-ups and SMEs, and consequently make the markets more competitive and more innovative.[24] Cloud computing also allows data-driven entrepreneurs to focus on creating and marketing innovative data-driven products without much concern about scaling computing and networking to fit demand.[25] A number of consulting companies have forecast tremendous growth in the public cloud computing market, particularly in the field of SaaS, in the next decade (Ried, 2011).

Surveys by the cloud computing technology provider VMware (2011) confirm that i) increasing business agility and ii) decreasing ICT investment costs are the main

motivations for business adoption of cloud computing. Fifty-seven per cent of all respondents point to accelerating the execution of projects and improvement of the customer's experience as the most frequent reasons for cloud computing adoption, followed by rapid adaption to market opportunities (56%) and the ability to scale cost (55%). Recent figures on the adoption of cloud computing reveal significant cross-country variations however (Figure 3.11). In countries such as Finland, Israel, Italy, Sweden and Denmark, almost half of all businesses are already using cloud computing services. There is also large variation by business size, with larger enterprises (250 or more employees) more likely to use cloud computing. In the United Kingdom, for instance, 21% of all smaller enterprises (10 to 49 employees) are using cloud computing services, compared to 54% of all larger enterprises.

Figure 3.11. **Enterprises using cloud computing services by *employment size class*, 2014**

As a percentage of enterprises in each employment size class



Note: Data for Canada refer to the use of "software as a service", a subcategory of cloud computing services.

Source: Based on Eurostat, Information Society Statistics, and Statistics Canada, January 2015.

The use of cloud computing brings other possible benefits that could greatly facilitate the introduction of DDI. Some cloud computing platforms come with standardised interfaces that make it easier to bring several services together, or to interconnect with another smart service that is operating on the same or another cloud platform. As a result, it becomes possible to integrate these services and thereby develop new innovative services. As manufacturers or operators of the smart device will not provide all services, a new market may emerge for companies that would offer to integrate the data from various devices into one package. A home management service may bring together data from sensors and actuators for lights, energy, temperature or movement with other types of sensors and devices, and so provide an integrated overview of all home services. Based on the data collected from many homes and other sources such as weather forecasting, the system may be able to optimise energy consumption. Such cloud-based services could be effective on a very wide scale, with large numbers of customers and large data sets involved.

Despite its widely recognised benefits, there remain significant issues limiting adoption of cloud computing. Privacy and security are among the two most pressing issues, which are discussed further in Chapter 5 of this volume. Another major challenge is the lack of appropriate standards and the potential for vendor lock-in due to the use of proprietary solutions (OECD, 2014d). According to recent surveys among potential users

of cloud computing, a lack of standards and the lack of widespread adoption of existing standards are seen as two of the biggest challenges. The lack of open standards is a key problem mainly in the area of PaaS. In this service model, application programming interfaces (APIs) are generally proprietary. Applications developed for one platform typically cannot be easily migrated to another cloud host. While data or infrastructure components that enable cloud computing (e.g. virtual machines) can currently be ported from selected providers to other providers, the process requires an interim step of manually moving the data, software, and components to a non-cloud platform and/or conversion from one proprietary format to another. As a consequence, once an organisation has chosen a PaaS cloud provider, it is – at least at the current stage – locked in (see Chapter 2).

3.3. From informing to driving decision-making

The exponential growth in the data generated and collected, combined with the pervasive power of data analytics, has led to a paradigm shift in the ways knowledge is created and – in particular – decisions are made. These two moments, namely when data are transformed into knowledge (gaining insights) and then used for decision making (taking action), are when the social and economic value of data is mainly reaped. The decision-making phase seems to be the most important one for businesses. According to a survey by the Economist Intelligence Unit (2012), for example, almost 60% of business leaders use big data for decision support and almost 30% for decision automation. This is echoed in estimates by Brynjolfsson, Hitt and Kim (2011), which suggest that the output and productivity of firms that adopt data-driven decision making are 5% to 6% higher than would be expected from their other investments in and use of information technology (see Chapter 1). This section highlights:

1. How value is created when knowledge is extracted from data.
2. How that knowledge is then used for data-driven decision making. Here, two major trends are highlighted: i) decision making is increasingly based on real-time experiments, and ii) it is automated.

Gaining insights: From data to information to knowledge

To understand the value creation process through data analytics, it helps to see data, information and knowledge as different but interrelated concepts. Information is often conveyed through data, while knowledge is typically gained through the assimilation of information. The boundaries between data, information and knowledge may seem extremely fuzzy sometimes, which explains why these concepts are often used as synonyms in media and literature (see Hess and Ostrom, 2007; Daniel Bell, cited in Cleveland, 1982).[26] However, separating these concepts is important to better understand data-driven value creation. A clearer distinction can also help explain certain paradoxes – for example, why one can have a lot of data, but not be able to extract value from them when not equipped with the appropriate analytic capacities (OECD, 2013b; Ubaldi, 2013). Similarly, one can have a lot of information, but not be able to gain knowledge from it – a phenomenon nowadays better known as "information overload" (see Speier et al., 2007)[27] and which Nobel prize-winning economist Herbert Simon described with the words: "a wealth of information creates a poverty of attention" (Shapiro and Varian, 1999). This section discusses the three main functions through which data analytics today is used to gain insights: i) extracting information from unstructured data; ii) real-time monitoring; and iii) inference and prediction. It is interesting to note here that

the first two functions are related to two of the three Vs which many see as the key characteristics of big data: variety and velocity (see Glossary). The first V (volume) refers to the exponential growth in data generated and collected, already discussed in the previous section.

Extracting information from unstructured data

Data analytics today has attracted a lot of attention due to its capacity to analyse in particular unstructured data – that is, data that lack a predefined data model (i.e. an abstract representation of "real world" objects and phenomena) (see Hoberman, 2010).[28] Data are considered structured if they are based on such a predefined model. These data models are needed for data processing and can be explicit, as in the case of a SQL database where the data model is reflected in the structure of the database's tables and their inter-linkages. The data model can also be implicit, as in the case of structured web content – or of web logs, where the underlying (implicit) model can be made explicit at relatively low cost. As they do not have an explicit but implicit model, these types of data are often referred to as semi-structured data. Semi-structured data can also refer to data without an explicit data model but to which are attached semantic elements such as tags that highlight the structure within the data. In contrast, with unstructured data, model can only be extracted at significant cost. Typical examples include text-heavy data sets such as text documents and emails, as well as multimedia content such as videos, images and audio streams.

Unstructured data are by far the most frequent type of data, and thus provide the greatest potential for data analytics today. According to a survey of data management professionals by Russom (2007), less than half of the total data stored in businesses is structured. The remaining data are either unstructured (31%) or semi-structured (21%). The author admits, however, that the real share of unstructured (including semi-structured) data could be much higher, as only data management professionals dealing mostly with structured data and rarely with unstructured data were surveyed. Older estimates suggest that the share of unstructured data could be as high as 80% to 85% (see Shilakes and Tylman, 1998).[29] A recent study by IDC (2012b) estimates that not even 5% of the "digital universe" is tagged, and thus can be considered structured or semi-structured data.

However, the difference between structured, semi-structured and unstructured data is becoming less important in the long run, since with growing computing capacities data analytics is increasingly able to automatically *extract* the information embedded in unstructured data. In the past, extracting that information was labour-intensive. The potential of data analytics for automating the processing of unstructured data sets can be illustrated via the evolution of search engines. Web search providers such as Yahoo! initially started with highly structured web directories edited by people. These services could not be scaled up as online content increased. Search providers had to introduce search engines that automatically crawled through "unstructured" web content, using links to extract even more information about the relevance of the content.[30] Yahoo! only introduced web crawling as the primary source of its search results in 2002. By then, Google had been using its search engine (based on its PageRank algorithm) for five years, and its market share in search had grown to more than 80% in 2012.[31] (See Watters, 2012 for a comparison of Yahoo! and Google in terms of structured vs. unstructured data.)

A series of technologies have further increased the capacity of data analytics to *process* unstructured data. Optical character recognition (OCR), for example, can

transform images of text into machine-encoded text, which then can be interpreted by software, such as for example when indexed for search services such as via Google Books. Natural language processing (NLP), another example, can then be used for tagging or for extracting relevant communication patterns and even emotional patterns. Twitter, for example, has been discussed as a potential (unstructured) data source for analysing and even predicting the "emotional roller coaster" and its impact on the ups and downs of stock markets (Grossman, 2010; *MIT Technology Review*, 2010). Other examples include applications based on face recognition, which – powered by machine-learning algorithms – are able to recognise individuals from images and even video streams. Facebook, for example, is known for using face recognition algorithms to automatically identify and tag its users out of user-provided images (Andrade, Martin and Monteleone, 2013).

Real-time monitoring and tracking

The speed at which data are collected, processed and analysed is often also highlighted as one of the key benefits of data analytics today. The collection and analysis of data in (near to) real time has empowered organisations to base decisions on "close-to-market" evidence. For businesses, this means reduction of time to market and first- or early-mover advantages. For governments, it can mean real-time evidence-based policy making (Reimsbach-Kounatze, 2015).[32] For example, policy analysts have come to use readily available data to make real-time "nowcasts", ranging from purchases of autos to flu epidemics to employment/unemployment trends, in order to improve the quality of policy and business decisions (Choi and Varian, 2009; Carrière-Swallow and Labbé, 2013). The Billion Price Project (BPP), launched at MIT and spun off to a firm called PriceStats, collects more than half a million prices on goods (not services) a day by "scraping the web". Its primary benefit is its capacity to provide real-time price statistics that are timelier than official statistics. In September 2008, for example, when Lehman Brothers collapsed, the BPP showed a decline in prices that was not picked up until November by the official Consumer Price Index (Surowiecki, 2011). Data analytics is also used for security purposes, such as real-time monitoring of information systems and networks to identify malware and cyberattack patterns. The security company ipTrust, for instance, computes and assigns reputation scores to IP addresses in real time to identify traffic patterns from bot-infected machines (Harris, 2011).

Inference and prediction: the new power of machine learning

Data analytics enables the "discovery" of information even if there was no prior record of such information. Such information can be derived in particular, as indicated earlier, by "mining" available data for patterns and correlations. As the volume and variety of available data sets increases, so does the ability to derive further information from these data, notably when they are linked. In particular, personal information can be "inferred" from several pieces of seemingly anonymous or non-personal data (see Chapter 5 of this volume). As the need for data analytics becomes more focused on real-time insights rather than historical and periodical information, the market demands for data analytics change as well, leading to higher demand for advanced specialised data analytic services. In addition, it is becoming increasingly important not only to generate the best actionable output, but also to present it in such a way that it is aligned with the business process that it strives to support, in order to establish competitive differentiation (Dumbill, 2011). For the next couple of years it is expected that most of the value added of data will come from advanced analytical techniques, in particular predictive analytics,

simulations, scenario development and advanced data visualisations, many of which are based on advanced use of machine learning (Russom, 2011).

Machine or statistical learning, as mentioned above, is based on the use of algorithms that allow computers to "learn" from data. Having analysed similar situations, computers can apply this analysis to infer and predict a present and a future situation. To make this work, machine learning uses many techniques that are also used in data analytics – for example, a large patient data set can help determine correlations for illnesses. Although machine learning involves such techniques, it is sometimes viewed as different from data analytics, which often attempts to describe the current situation and to find new and unknown correlations in the data. But the distinction is blurring, as machine learning relies on common techniques such as statistical and regression analysis of data to determine future actions in new situations, while data analytics increasingly relies on (unsupervised) learning algorithms for inference and prediction – for instance, via cluster analysis (Hastie, Tibshirani and Friedman, 2011; James et al., 2013). For an historical perspective of machine learning, see Box 3.3.

Web services are notably an area where machine learning is very important. Many of the modern tools and techniques that have become available were developed for web-based services. Search engines are large-scale users of machine-learning technologies, which is not surprising given their relation to translation and speech recognition. Related to this field are the recommendation engines that power services such as Amazon, Deezer, Spotify and Netflix. These services use machine learning to predict the goods that best fit a user's taste. In order to determine this, they use data on the ratings given by the users, for instance to music, as well as information on how they used the service – for example, skipping a song or stopping a movie halfway through and not returning to it. These algorithms are essential to the success of the service: research has shown that consumers will not make a decision when faced with too many options. Machine learning reduces the stress associated with choice (*The Economist*, 2010). Netflix went as far as organising a contest where it awarded winners USD 1 million for the best predictive algorithm. Netflix (2012) tests the algorithms by performing A|B tests,[33] where different algorithms are pitched against each other and their success is measured.

Box 3.3. **Machine learning: An historical perspective**

Translation and speech recognition was one of the first areas where artificial intelligence (AI) was applied. The traditional approach was to describe all the rules related to a language in the software such as the grammar, but also the meaning of words in context. The complexity came from teaching the computer rules to determine the difference among the meanings of one word, such as for instance right as correct, right as a direction, and other meanings of the word. The academic work was mostly performed by linguists, who benefited from an ever better knowledge of language and its rules as a result. However, the computer systems failed to be practical. The alternative approach, statistical analysis of data to derive probabilities, had been discussed in the late 1950s and 1960s. This approach, however, found opposition from noted academics such as Noam Chomsky, who wrote: "we are forced to conclude that grammar is autonomous and independent of meaning and that probabilistic models give no insight into the basic problems of syntactic structure" (quoted in Young, 2010). As a result, the statistical approach was not given full academic attention for some decades.

Box 3.3. **Machine learning: An historical perspective (cont.)**

In 1976, a seminal paper titled "Continuous speech recognition by statistical methods" was published by Frederick Jelinek of IBM. Jelinek (1997) approached the problem of speech recognition not as a linguistic problem, but as a mathematical problem on a par with signals analysis in fixed and wireless networks. This was the start of the resurgence of statistical methods. What made his approach unique was that it relied on statistical analysis of speech and language and not on complex rule models of language. This required the training of the system with many examples of the language. From so-called n-grams (trigrams), combinations of generally three words that were commonly together were derived, and statistics were used to best fit the matching words or pronunciation. The results were significantly better systems compared to earlier rule-based systems, despite the fact that the system was not "knowing" why the result was better.

Today the statistical approach is the basis of speech recognition and translation, such as Siri of Apple and Google Now, and online translation tools offered by Google, Microsoft and Yahoo. Systems are trained by feeding them large corpora of texts, such as subtitled television programmes and the official translations into the official languages of the Canadian Parliament, European Union and United Nations, but also by web pages and scanned books. The results, though not perfect, are often usable. The application continuously adds to the systems by scanning more and more data and by analysing user-provided corrections to texts.

But machine learning is not used solely by Internet firms. In health care, for example, data collected on patients are recorded by imaging and other sensors. Data on the environment in both the health care facilities and the patients' environment can be of relevance as well. Researchers are therefore looking into machine-learning algorithms to better detect conditions and at the same time cut back the number of false positives and negatives. In Chinese Taipei, researchers report that a system using machine learning delivered better results in avoiding false positives while determining three metabolic diseases in newborns. The system was trained on the data of close to 350 000 newborns, which had been collected and tested in prior years (Chen, 2013).

Machine learning is used in industrial applications as well.[34] One of the earlier examples was use in steel mills, where rollers of steel had to apply a controlled force on a hot piece of steel to achieve a particular thickness (Tresp, 2010). The traditional model used an approach based on analytical formulas. However, many effects were non-linear and therefore difficult to model and predict. Using machine learning, the error rate was significantly reduced.

Human decision making: Towards a business culture of data-driven experiments

The ubiquity of data generation and collection has enabled organisations to base their decision-making process on data even more than in the past. Two major trends deserve to be highlighted here: i) human decision making is increasingly based on rapid data-driven experiments; and ii) crowdsourcing –"the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community" (Merriam-Webster, 2014) – has been made further affordable thanks to the increased capacity to extract information from unstructured data from the Internet, and to share data with other analysts.

In business, for example, an increasing number of companies are crowdsourcing and analysing data as diverse as online, social media and sensor data to improve the design

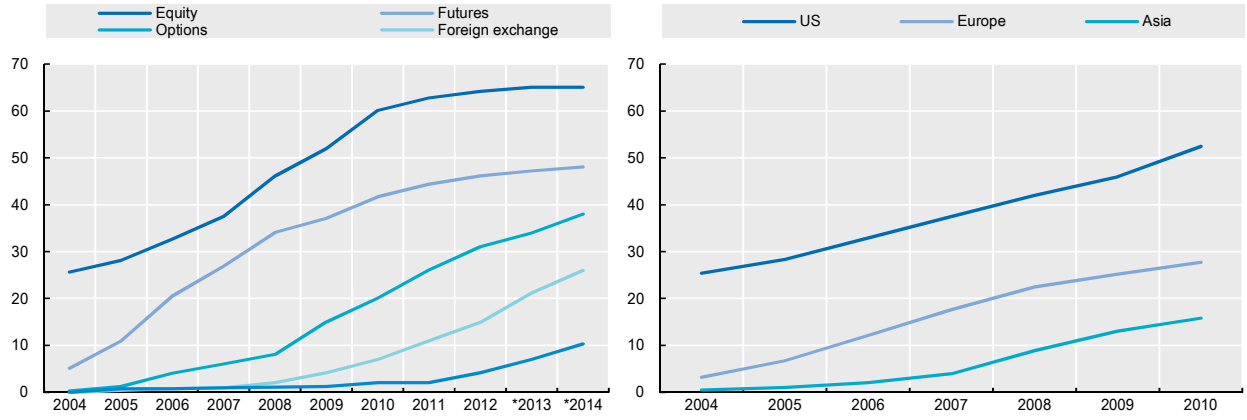
and quality of their products early in the design phase. They are also analysing these data sources to identify product-related problems to swiftly recall the products if necessary. The rapid analysis of these data sources enables firms to explore different options during product (re-)design and to reduce their opportunity costs and their investment risks. The online payment platform WePay, for instance, designs its web services based on A|B testing. For two months, users are randomly assigned a testing site. The outcome is then measured to determine whether the change in design led to statistically relevant improvements (Christian, 2012). Another example is John Deere, the agriculture equipment manufacturer, which provides farmers with a wide range of agricultural data that enable them to optimise agricultural production by experimenting with the selection of crops, and where and when to plant and plough the crops (Big Data Startups, 2013).

The use of data analytics in decision-making processes described above points to a shift in the way decisions are made in data-driven organisations. Decision makers do not necessarily need to understand the phenomenon before they act on it. In other words: first comes the analytical fact, then the action, and last, if at all, the understanding. For example, a company such as Wal-Mart Stores may change the product placement in its stores based on correlations without the need to know *why* the change will have a positive impact on its revenue. As Anderson (2008) explains: "Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity." Anderson has even gone as far as to challenge the usefulness of models in an age of massive data sets, arguing that with large enough data sets, machines can detect complex patterns and relationships that are invisible to researchers. The data deluge, he concludes, makes the scientific method obsolete, because correlation is enough (Anderson, 2008; Bollier, 2010). This has opened the door to increasing numbers of applications for decision automation (autonomous machines and systems), while raising "key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorisation of reality" (Boyd and Crawford, 2011; see Chapter 7 of this volume).

Autonomous machines and machine decision making

Data-driven decision making does not stop with the human decision maker. In fact, one of the largest impacts of DDI on (labour) productivity can be expected to come from decision automation, due to "smart" applications that are "able to learn from previous situations and to communicate the results of these situations to other devices and users" (OECD, 2013b). These applications are powered by machine-learning algorithms that are getting more and more powerful. They can perform an increasing number of tasks that required human intervention in the past. Google's driverless car is an illustration of the potential of smart applications. It is based on the collection of data from all the sensors connected to the car (including video cameras and radar systems), and combines it with data from Google Maps and Google Street View (for data on landmarks, traffic signs and lights). Another example is algorithmic trading systems (ATS) that can autonomously decide what stock to trade, when to trade it, and at what price. In the United States, ATS are estimated to account for more than half of all trades today (Figure 3.12).

Figure 3.12. **Algorithmic trading as a share of total trading**



Note: 2013-14 based on estimates.

Source: Based on *The Economist*, 2012.

Autonomous machines are seen as having great potential in logistics, manufacturing and agriculture. In manufacturing, robots have traditionally been used mostly where their speed, precision, dexterity and ability to work in hazardous conditions are valued. This is radically changing because of sensors, machine learning and cloud computing. Some modern factories, such as the Philips shaver factory in Drachten in the Netherlands, are almost fully robotic (Markoff, 2012). It employs only one-tenth of the workforce employed in its factory in China that makes the same shavers (see Chapter 6 for further discussion on the skills and employment implications of autonomous machines and machine decision making).

The limits of data-driven decision making

The use of data and analytics does not come without limitations, which given the current "big data" hype are even more important to acknowledge. There are considerable risks that the underlying data and analytic algorithms could lead to unexpected false results. The risks are higher where decision making is automated – as illustrated by the case of the Knight Capital Group, which lost USD 440 million in 2012, most of it in less than an hour, because it's ATS behaved unexpectedly (Mehta, 2012). Users should be aware of these limitations; otherwise they may (unintentionally) cause social and economic harm (costs), to themselves as well as to third parties. The risk of social and economic costs to third parties (including individuals) raises important questions related to the attribution of responsibility for inappropriate decisions.

That risk also raises the question of the extent to which the risk-based approach to security and privacy discussed in Chapter 5 allows taking into account all potential (negative) externalities. At times the incentives for the data and analytic user (i.e. the data controller) to minimise the risks to third parties may indeed be low. This is typically the case where the third parties will bear the main share of the social and economic costs of the data controller's action. Accordingly, there should be a careful examination both of the appropriateness of fully automated decision making, and of the need for human intervention in areas where the potential harm of such decisions may be significant (e.g. harm to the life and well-being of individuals, denial of financial or social rights).

Thought must also be given to increasing the transparency of the processes and algorithms underlying these automated decisions (i.e. algorithmic transparency),[35] while preserving proprietary intellectual property rights (IPRs) including in particular trade secrets, which some businesses would consider the "secret sauce" of their business operations (see OECD, 2015b).

The following types of errors are discussed further below: i) data errors; ii) errors that come with inappropriate use of data and analytics; and iii) errors caused by unexpected changes in the environment from which data are collected (i.e. the data environment). The latter issue is particularly relevant for decision automation.

Poor-quality data

The information that can be extracted from data depends on the quality of the data. Poor-quality data will therefore almost always lead to poor results ("garbage in, garbage out"). Therefore, data cleaning (or scrubbing) is often emphasised as an important step before the data can be analysed. And this often involves significant costs, as it can account for 50% to 80% of a data analyst's time together with the actual data collection (Lohr, 2014). As highlighted in Chapter 4, information is context dependent, and as a result data quality will typically depend on the intended use of the data: data that are of good quality for certain applications can thus be of poor quality for other applications (Lohr, 2014). The OECD (2011c) *Quality Framework and Guidelines for OECD Statistical Activities* therefore defines data quality as "fitness for use" in terms of user needs: "If data is accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data". The OECD (2013c) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) also provides a number of criteria for data quality in the context of privacy protection. The Recommendation states that "personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date".

Inappropriate use of data and analytics

As highlighted above, some have suggested that with big data, decision makers could base their actions only on analytical facts without the need to understand the phenomenon on which they are acting. As correlation would be enough with big data, scientific methods and theories would be less important. While it is true that analytics can be effective in detecting correlations in "big data", especially those that would not be visible with smaller-sized volumes of data, it is also widely accepted among practitioners that data analysis itself relies on rigorous scientific methods in order to produce appropriate results.

The rigour starts with how the quality of the data is assessed and assured. But even if data are of good quality, data analytics can still lead to wrong results if the data used are irrelevant and do not fit the business or scientific questions they are supposed to answer (see section above). Experts recognise that it is often too tempting to think that with big data one has sufficient information to answer almost every question and to neglect data biases that could lead to false conclusions, because correlations can often appear statistically significant even if there is no causal relationship. Marcus and Davis (2014) give the illustration of big data analysis revealing a strong correlation of the United States

murder rate with the market share of Internet Explorer from 2006 to 2011. Obviously, any causal relationship between the two variables is spurious.

The risk of inappropriate use of data and analytics underlines the need for high skills in data analysis, and challenges the belief that everyone and every organisation today is in a position to apply data analytics appropriately (see Chapter 6 on the skills implications of DDI). As O'Neil (2013a) argues, the simplicity of applying machine-learning algorithms today thanks to software improvements makes it easy for non-experts to believe in software-generated answers that may not correspond to reality. Furthermore, the need for understanding causal relationships means that sufficient domain-specific knowledge is necessary to apply data and analytics effectively. Obviously, the availability of high skills in data analysis and the rigorous use of data and analytics do not prevent data and analytics from being wrongly used intentionally for economic, political or other advantages. Literature is full of cases where (e.g.) sophisticated econometric models have been used to lie with data. O'Neil (2013b) discusses examples.

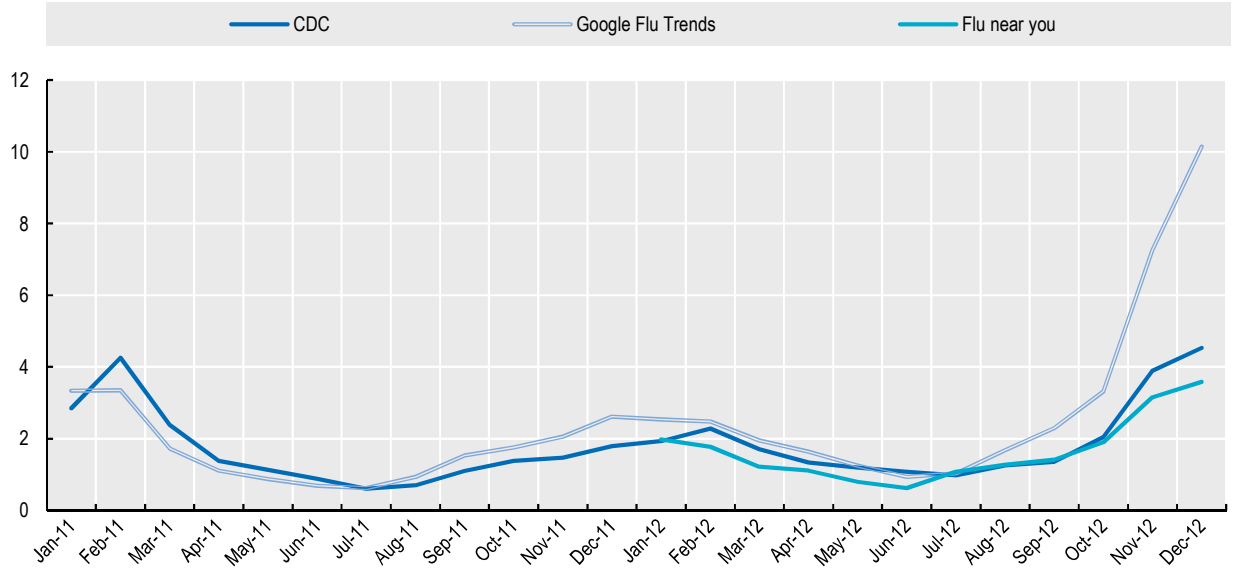
The changing data environment

Even when the data and the analytics are perfectly used initially, this does not mean that they will always deliver the right results. Data analytics, in particular when used for decision automation, can sometimes be easily "gamed" once the factors affecting the underlying algorithms have been understood – for example, through reverse engineering. Marcus and Davis (2014) for example present the case where academic essay evaluation analytics that relied on measures like sentence length and word sophistication to determine typical scores given by human graders, were gamed by students who suddenly started "writing long sentences and using obscure words, rather than learning how to actually formulate and write clear, coherent text". More popular examples (with business implications) are techniques known as "Google bombing" and "spamdexing", where users are adjusting Internet content, links and sites to artificially elevate website search placement in search engines (Segal, 2011; Marcus and Davis, 2014).

Data analytics does not need to be intentionally gamed to lead to wrong results. Often it is simply not sufficiently robust to address unexpected changes in the data environment. This is because data analytics users (including the developers of autonomous systems) cannot envision all eventualities that could affect the functioning of their analytic algorithms and software, in particular when they are used in a dynamic environment. In other words, data analytics is not perfect and some environments are more challenging than others. The case of the Knight Capital Group, which lost USD 440 million in financial markets in 2012 due to some unexpected behaviour from its trading algorithm, was already mentioned above. A more recent example is Google Flu Trends, which is based on Google Insights for Search and provides statistics on the regional and time-based popularity of specific keywords that correlate with flu infections.[36] Google Flu Trends has been used by researchers and citizens as a means to accurately estimate flu infection trends, and this at faster rates than the statistics provided by the Centers for Disease Control and Prevention (CDC). However, in January 2013, Google Flu Trends drastically overestimated flu infection rates in the United States (Figure 3.13). Experts assessed that this was due to "widespread media coverage of [that] year's severe US flu season", which triggered an additional wave of flu-related searches by people unaffected by flu (Butler, 2013).

Figure 3.13. **Fever estimations in the United States, January 2011-December 2012**

Estimated % of population with influenza-like illness, monthly average



Source: Based on Butler, 2013.

These incidents, intentioned or not, are caused by the dynamic nature of the data environment. The assumptions underlying many data analytics applications may change over time, either because users suddenly change their behaviour in unexpected ways as presented above (essay evaluation analytics) or because new behavioural patterns emerge out of the complexity of the data environment (algorithmic trading). As Lazer et al. (2014) further explain, one major cause of the failures (such as in the case of Google Flu Trends) may have been that the Internet constantly changes, and as a result the Google search engine itself constantly changes. Patterns in the data collected are therefore hardly robust over time.

3.4. Key findings and policy conclusions

This chapter has highlighted the key enablers of data-driven innovation, the understanding of which is crucial for governments to assess the degree of readiness of their economies to take advantage of DDI. Economies in which these enablers are more prevalent are expected to be in a better position to reap the benefits of DDI. This does not mean that all factors need to be fully developed in order to realise those benefits. As shown in Chapter 2 of this volume, the global nature of the data ecosystem allows countries to profit from DDI through data- and analytics-related goods and services produced elsewhere. However, it can be assumed that countries with enhanced capacities to supply *and* use data and analytics will be in the best position.

A fast and open Internet (including the Internet of Things) is the most fundamental condition for DDI. In particular:

1. Mobile broadband enables mobile devices (many of which are smart devices enabled by M2M and sensors) to be used for DDI, including in remote and less developed areas where DDI could bring much needed (regional) growth (e.g. DDI

in agriculture). However, while in Finland, Australia, Japan, Sweden, Denmark and Korea mobile penetration rates exceeded 100%, they are still at 40% or less in Portugal, Greece, Chile, Turkey, Hungary and Mexico.

2. The functioning of co-location and backhaul markets is key for the local deployment of data-driven services. Analysis of the share of the most popular local content sites hosted domestically suggests that the local market for hosting and co-location is not functioning efficiently in countries with a low proportion of their most popular local content sites hosted domestically. Underlying reasons may differ vastly from country to country and may deserve for follow-up studies.

3. There are regulatory barriers preventing effective deployment of some M2M-based mobile applications. In particular, large-scale M2M users such as car manufacturers who need to control their own devices with their own SIM cards cannot do so in many countries, as it would make a car manufacturer the equivalent of a mobile operator. Removing regulatory barriers to entry in the mobile market would allow the million-device customer to become independent of the mobile network and to further competition.

4. Barriers to the open Internet, whether legitimate or not, can limit the effects of DDI. Some of these barriers may be technical, such as IP package filtering, or regulatory, such as "data localisation" requirements, and they may be the results of business practices and government policies. Some of these have a legal basis such as privacy and security (see Chapter 5) as well as the protection of trade secrets and copyright (see OECD, 2015b). However, these barriers can have an adverse impact on DDI – for example, if they limit trade and competition (see Chapter 2). Governments looking to promote DDI in their countries should take the OECD (2011b) *Council Recommendation on Principles for Internet Policy Making* further into consideration as well as ongoing OECD work to develop a better understanding of the characteristics, and the social and economic impact of the open Internet.

Data analytics and super computing power are complementary resources needed for the use of "big data". Access to these resources is therefore critical for realising the potential of DDI. However, there are two important issues:

1. Lack of interoperability and the risk of vendor lock-in are two major concerns potential cloud computing users have that may warrant policy makers' attention. The lack of open standards is mainly a huge problem in the area of PaaS. Initiatives are under way to address this issue, covering the full spectrum from infrastructure standards – such as virtualisation formats and open APIs for management to standards for web applications and services and data linkage, but also privacy, security and identity management.

2. Access to and effective use of data analytics can be affected by IPR, in two ways. First, data analytics (including its algorithms) can be protected by software patents or copyright, which under some conditions can limit access and the range of applications. Second, in the special case of text mining, the use of data analytics can be restricted due to copyright, even where scientists may have legal access to scientific publications. While the first issue may not always pose a serious problem to the data analytics community, the latter is still subject to controversial debates between the scientific community and the publishers of scientific journals.

Finally, analysis of value creation mechanisms shows that:

1. Data analytics leads to new ways of decision making, in particular through low cost and rapid experiments (often based on correlations and A|B testing), as well as through use of autonomous machines and systems (based on machine-learning algorithms) that are able to learn from previous situations and to (autonomously) improve decision making.
2. However, there are serious risks that the use of data and analytics may lead to inappropriate results. Strong skills are thus needed in data analysis and domain-specific knowledge, as discussed in Chapter 6. This challenges the "democratisation" of data analytics, according to which everyone and every organisation can use data and analytics appropriately. The risks are elevated when analytics are used for decision automation in dynamic environments, in which case the environments need to be properly understood as well. Likewise, careful examination of the appropriateness of fully automated decision making and of the need for algorithmic transparency and human intervention is critical in areas where the potential harm of such decisions may be significant (e.g. harm to the life and well-being of individuals, denial of financial or social rights).

Notes

1 It is estimated that the Babylonian census, introduced in 1800 BC, was the first practice of systematically counting and recording people and commodities for taxation and other purposes. See www.wolframalpha.com/docs/timeline.

2 Luhn (1958) introduces the concept of business intelligence, citing the following Webster's Dictionary definition of intelligence: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal". He further defines business as "a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera."

3 As Mayer-Schönberger and Cukier (2013) explain: "To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analyzed".

4 It has been argued however that in some cases, additional measures guaranteeing the delivery of time-sensitive data may be needed (e.g. quality of service).

5 For that study's analysis, "the generic top-level domains were omitted from the list, as there is no reliable public data as to where the domains are registered. Out of the one million top sites, 946 700 were scanned, 474 000 were generic top-level domains, 40 000 had no identifiable host country, and 3 700 had no identifiable domain, just an IP-address. The remaining 429 000 domains were analysed and their hosting country identified. For each country the percentage of domains hosted in the country were [sic] identified". See also Royal Pingdom blog, available at http://royal.pingdom.com/2012/06/27/tiny-percentage-of-world-top-1-million-sites-hosted-africa/.

6 As discussed in the study, there are caveats that need to be highlighted – for example, the high share of generic top-level domains hosted in the United States for historical reasons. The ccTLD .us is also a valid top-level domain in that country, but it is very lightly used. … There are some further caveats with the data. In some cases there may be a national and an international site for the content. For example, it might be the case that a newspaper has a site hosted in the country, for all web requests coming from the country and an international site located close to where the countries [sic] diaspora lives. The local site will likely not show up as the query was run from Sweden. Similarly, some of the largest sites in the world use content delivery networks (CDNs) to distribute their data. These sites show as hosted outside the country, though for visitors in country, they may be local".

7 This comes as no surprise considering the importance of reliable energy supply for the operation of data centres (Reimsbach-Kounatze, 2009).

8 Mashups or mash-ups are web applications that use and combine content from different sources, including but not limited to web documents such as web pages and multimedia content; data such as cartographic and geographic data; application converters; and communication and visualisation tools.

9 Today a standard smartphone, for example, contains the following sensors besides microphones and video sensors: (i) accelerometer – measures magnitude and direction of acceleration, (ii) global positioning system (GPS) – measures location based on the position of satellites, (iii) gyroscope – measures orientation of a device, (iv) barometer – measures air pressure, which is also used to measure vertical movement, and (v) magnetometer (compass) – measures device orientation.

10 These sensors can be regarded as "the interface between the physical world and the world of electrical devices, such as computers" as they measure multiple physical properties. Examples include electronic sensors, biosensors, and chemical sensors (see Wilson, J. (2008), Sensor Technology Handbook, Newnes/Elsevier, Oxford). The counterpart is represented by actuators that function the other way round, i.e. whose tasks consist in converting the electrical signal into a physical phenomenon (e.g. displays for quantities measures by sensors such as speedometers, temperature reading for thermostats, but also those that control the motion of a machine).

11 "The MIT License is a permissive license that is short and to the point. It lets people do anything they want with your code as long as they provide attribution back to you and don't hold you liable. jQuery and Rails use the MIT License." (See http://choosealicense.com/.)

12 The BSD License is "a permissive license that comes in two variants, the BSD 2-Clause and BSD 3-Clause. Both have very minute differences to the MIT license." (See http://choosealicense.com/licenses/.)

13 "The Apache License is a permissive license similar to the MIT License, but also provides an express grant of patent rights from contributors to users. Apache, SVN, and NuGet use the Apache License." (See http://choosealicense.com/.)

14 "The GPL (V2 or V3) is a copyleft license that requires anyone who distributes your code or a derivative work to make the source available under the same terms. V3 is similar to V2, but further restricts use in hardware that forbids software alterations. Linux, Git, and WordPress use the GPL." (See http://choosealicense.com/.)

15 A well-known example is R, a GPL-licensed open source environment for statistical analysis, which is increasingly used as an alternative to commercial packages such as SPSS and SAS (see section below). Another example is the library scikit-learn, which provides a set of data analytics and machine-learning algorithms for the programming language Python, and is provided under the BSD License. It was developed during a Google Summer of Code project as a third party extension to a separately developed Python project, SciPy, a BSD-licensed open source ecosystem for scientific and technical computing.

16 As Paul explains: "Many companies in technical fields attempt to collect as many broad patents as they can so that they will have ammunition with which to retaliate when they are faced with patent infringement lawsuits." For more on IP strategies see OECD (2015b).

17 IBM is offering its Hadoop solution through InfoSphere BigInsights. BigInsights augments Hadoop with a variety of features, including textual analysis tools that help identify entities such as people, addresses and telephone numbers (Dumbill, 2012a).

18 Oracle provides its Big Data Appliance as a combination of open source and proprietary solutions for enterprises' big data requirements. It includes, among others, the Oracle Big Data Connectors that allows customers to use Oracle's data warehouse

and analytics technologies together with Hadoop; the Oracle R Connector, which allows the use of Hadoop with R; and the Oracle NoSQL Database, which is based on Oracle Berkeley DB, a high-performance embedded database.

19 In 2011, Microsoft started integrating Hadoop in Windows Azure, Microsoft's cloud computing platform, and one year later in Microsoft Server. It is providing Hadoop Connectors to integrate Hadoop with Microsoft's SQL Server and Parallel Data Warehouse (Microsoft, 2011).

20 In 2012, SAP announced its roadmap to integrate Hadoop with its real-time data platform SAP HANA and SAP Sybase IQ.

21 Surveys on the use of data analytics software are also confirming these results. A survey by KDnuggets, for example, suggests that RapidAnalytics (free edition), R, Excel, Weka/Pentaho, and Python were the top five data analytics tools used in 2013. Although all except Excel are free or open source tools, the authors of the survey conclude that commercial and free/open source software are used almost equally among the surveyed data analysts.

22 Four of the top five packages used were open source, including RapidMiner (free edition), R, Weka/Pentaho, and the combination of Python tools numpy, scipy and panda.

23 Sometimes, clouds are also classified into private, public, and hybrid clouds according to their ownership and control of management of the clouds.

24 Due to economies of scale, cloud computing providers have much lower operating costs than companies running their own IT infrastructure, which they can pass on to their customers.

25 Big data solutions are typically provided in three forms: software-only, as a software-hardware appliance, or cloud-based (Dumbill, 2012b). Choices among these will depend, among other things, on issues related to data locality, human resources, and privacy and other regulations. Hybrid solutions (e.g. using on-demand cloud resources to supplement in-house deployments) are also frequent.

26 According to Hess and Ostrom (2007), "knowledge […] refers to all intelligible ideas, information, and data in whatever form in which it is expressed or obtained". Daniel Bell defines information as "data processing in the broadest sense" and knowledge as "an organized set of statements of facts or ideas […] communicated to other".

27 As Speier et al. explain: "Information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have fairly limited cognitive processing capacity. Consequently, when information overload occurs, it is likely that a reduction in decision quality will occur."

28 According to Hoberman, "a data model is a wayfinding tool for both business and IT professionals, which uses a set of symbols and text to precisely explain a subset of real information to improve communication within the organization and thereby lead to a more flexible and stable application environment."

29 In health care, for example, health records and medical images are the dominant type of data, and they are sometimes stored as unstructured data. Estimates suggest that in the United States alone, 2.5 petabytes are stored away each year from mammograms.

30 See Watters (2012) for a comparison of Yahoo! and Google in terms of structured *vs.* unstructured data.

31 See http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4.

32 Real-time data can also be a source for real-time evidence for policy making. The Billion Price Project (BPP), for example, collects price information over the Internet to compute a daily online price index and estimate annual and monthly inflation. It is not only based on five times what the US government collects, but it is also cheaper, and is has a periodicity of days as opposed to months.

33 A|B testing is typically based on a sample that is split into two groups, an A group and a B group. While an existing strategy is applied to the (larger) A group, another, slightly changed strategy is applied to the other group. The outcome of both strategies is measured to determine whether the change in strategy led to statistically relevant improvements. Google, for example, regularly redirects a small fraction of its users to pages with slightly modified interfaces or search results to (A|B) test their reactions.

34 Now that sensor data are becoming more widely available in industrial applications, companies such as Siemens and General Electric are increasingly promoting machine-learning applications.

35 At the fourth meeting of the OECD Global Forum on the Knowledge Economy (GFKE) on "Data-driven Innovation for a Resilient Society", held 2-3 October 2014 in Tokyo (www.gfke2014.jp/), EPIC President, Marc Rotenberg, highlighted the need for "algorithmic transparency", which would make data processes that impact individuals public (see Annex of Chapter 1 of this volume on the highlights of the GFKE).

36 Google Trends now also include surveillance for a second disease, dengue.

References

Abedi, V. et al. (2012), "An automated framework for hypotheses generation using literature", *BioData Mining*, Vol. 5, No. 1.

Anderson, C. (2008), "The end of theory: The data deluge makes the scientific method obsolete", *Wired*, 23 June, www.wired.com/science/discoveries/magazine/16-07/pb_theory/, accessed 05 May 2015.

Andrade, N.N.G, A. Martin and S. Monteleone (2013), "'All the better to see you with, my dear': Facial recognition and privacy in online social networks", *IEEE Security & Privacy*, Vol. 11, No. 3, pp. 21-28, May/June.

Automotive Sensors Conference (2015), Conference website, www.automotivesensors2015.com, accessed 21 October 2014.

Bekhuis, T. (2006), "Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy", *Biomed Digit Library*, http://dx.doi.org/10.1186/1742-5581-3-2 (accessed 13 August 2014).

Bertolucci, J. (2013), "IBM, universities team up to build data scientists", *InformationWeek*, 15 January, www.informationweek.com/big-data/big-data-analytics/ibm-universities-team-up-to-build-data-scientists/.

Big Data Startups (2013), "Walmart is making big data part of its DNA", www.bigdata-startups.com/BigData-startup/walmart-making-big-data-part-dna/, accessed 22 August 2014.

Bollier, D. (2010), *The Promise and Peril of Big Data*, Aspen Institute, Washington, DC.

Boyd, D. and K. Crawford (2011), "Six Provocations for Big Data", A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, 21 September, http://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431.

Brynjolfsson, E., L.M. Hitt and H.H. Kim (2011), "Strength in numbers: How does data-driven decision making affect firm performance?", 22 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

Butler, D. (2013), "When Google got flu wrong", *Nature*, 13 February, www.nature.com/news/when-google-got-flu-wrong-1.12413.

Carrière-Swallow, Y. and F. Labbé (2013), "Nowcasting with Google trends in an emerging market", *Journal of Forecasting,* Vol. 32, No. 4, pp. 289-98.

Chang, F. et al. (2006), "Bigtable: A distributed storage system for structured data", Google Research Publications, appeared in the proceedings of the Seventh Symposium on Operating System Design and Implementation (OSDI'06), November, http://research.google.com/archive/bigtable.html.

Chen, W.-H. et al. (2013), "Web-based newborn screening system for metabolic diseases: Machine learning versus clinicians", Journal of Medical Internet Research, www.jmir.org/2013/5/e98/.

Choi, H. and H. Varian (2009), "Predicting the present with Google trends", *Social Science Electronic Publishing*, 10 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302.

Christian, B. (2012), "The A|B Test: Inside the technology that's changing the rules of business", *Wired*, 25 April, www.wired.com/business/2012/04/ff_abtesting.

Cisco (2013), "Cisco Visual Networking Index: Forecast and methodology", 2012-2017, www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html, accessed 12 November 2013.

Cleveland, H. (1982), "Information as a resource", *The Futurist*, December, http://hbswk.hbs.edu/pdf/20000905cleveland.pdf, accessed 12 September 2013.

Davis, A.P. et al. (2013), "A CTD-Pfizer collaboration: Manual curation of 88 000 scientific articles text mined for drug-disease and drug-phenotype interactions", *Database*, http://dx.doi.org/10.1093/database/bat080 published online 28 November.

Dean, J. and S. Ghemawat (2004), "MapReduce: Simplified data processing on large clusters", Sixth Symposium on Operating System Design and Implementation (OSDI'04), San Francisco, December, http://research.google.com/archive/mapreduce.html.

Duda, R., P.E. Hart and D.G. Stork (2000), *Pattern Classification*, Second Edition, Wiley-Interscience, 9 November.

Dumbill, E. (2012a), "Big data market survey: Hadoop solutions", *O'Reilly Radar*, 19 January, http://radar.oreilly.com/2012/01/big-data-ecosystem.html.

Dumbill, E. (2012b), "What is big data? An introduction to the big data landscape", *O'Reilly Radar*, 11 January, http://radar.oreilly.com/2012/01/what-is-big-data.html.

Dumbill, E. (2011), "Five big data predictions for 2012", *O'Reilly Radar*, http://strata.oreilly.com/2011/12/5-big-data-predictions-2012.html.

Economist Intelligence Unit (2012), "The deciding factor: Big data & decision making", Economist Intelligence Unit commissioned by Capgemini, 4 June, http://www.capgemini.com/insights-and-resources/by-publication/the-deciding-factor-big-data-decision-making.

Ericsson (2010), "CEO to shareholders: 50 billion connections 2020", Ericsson press release, 13 April, www.ericsson.com/thecompany/press/releases/2010/04/1403231, accessed 12 January 2014.

Goetz, T. (2012), "How to spot the future", *Wired*, 24 April, www.wired.com/2012/04/ff_spotfuture/.

Grossman, L. (2010), "Twitter can predict the stock market", *Wired*, 19 October, www.wired.com/wiredscience/2010/10/twitter-crystal-ball/.

Gurulingappa, H. et al. (2013), "Automatic detection of adverse events to predict drug label changes using text and data mining techniques", *Pharmacoepidemiology and Drug Safety*, Vol. 22, No. 11, November, pp. 1189-94.

Harris, D. (2011), "Hadoop kills zombies too! Is there anything it can't solve?", *Gigaom*, 18 April, http://gigaom.com/cloud/hadoop-kills-zombies-too-is-there-anything-it-cant-solve/.

Hastie, T., R. Tibshirani and J. Friedman (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer, New York.

Hazard Owen, L. (2013), "Add $0.50 worth of sensors to your iPhone 5s and it'll be able to track your emotions", *Gigaom,* https://gigaom.com/2013/10/17/add-0-50-worth-of-sensors-to-your-iphone-5s-and-itll-be-able-to-track-your-emotions, accessed 21 October 2014.

Hess, C. and E. Ostrom (2007), *Understanding Knowledge as a Commons: From Theory to Practice,* MIT Press, Cambridge, Mass.

Hey, J. (2004), "The data, information, knowledge, wisdom chain: The metaphorical link", working paper, December, www.dataschemata.com/uploads/7/4/8/7/7487334/dikwchain.pdf, accessed 12 March 2013.

Hoberman, S. (2010), "How do you justify data modeling?", *Erwin Expert Blog*, 20 April, http://erwin.com/community/expert-blogs/how-do-you-justify-data-modeling.

Inmon, W.H. and C. Kelly (1992), *Rdb/VMS: Developing the Data Warehouse*, QED Publishing.

IDC (International Data Corporation) (2012a), "The Digital Universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East", EMC Digital Universe project, IDC *iView*, December, www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf, accessed 15 September 2013.

IDC (2012b), "Worldwide big data technology and services 2012-2015 forecast", IDC Market Analysis, March, www.idc.com/research/viewtoc.jsp?containerId=233485.

Inaudi, D. and A. del Grosso (2011), "Fiber optic sensors for structural control", www.roctest-group.com/sites/default/files/bibliography/pdf/c196.pdf, accessed 5 September 2011.

ITU (2014), "The world in 2014: ICT facts and figures", International Telecommunication Union, April, www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf.

Janert, P. (2010), *Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists*, O'Reilly Media.

James, G. et al. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Jelinek, F. (1997), *Statistical Methods for Speech Recognition*, MIT Press.

JISC (2012), "The value and benefits of text mining", JISC**,** www.jisc.ac.uk/sites/default/files/value-text-mining.pdf, accessed 14 June 2014.

Keen, P.G.W. (1978), *Decision Support Systems: An Organizational Perspective*, Addison-Wesley, Reading, Mass.

Lazer, D. et al. (2014), "The parable of Google flu: Traps in big data analysis", *Science*, Vol. 343, No. 14, March, http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

KDnuggets (2013), "What analytics, big data, data mining, data science software you used in the past 12 months for a real project?", KDnuggets, May,

www.kdnuggets.com/polls/2013/analytics-big-data-mining-data-science-software.html.

Kundra, V. (2011), "Federal cloud computing strategy", US Chief Information Officers Council, www.cio.gov/documents/federal-cloud-computing-strategy.pdf, accessed 7 October 2013.

Leipzig, J. and X. Li (2011), *Data Mashups in R: A Case Study in Real-World Data Analysis*, O'Reilly Media.

Lohr, S. (2014), "For big-data scientists, 'janitor work' is key hurdle to insights", *New York Times*, 17 August, www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Lok, C. (2010), "Literature mining: Speed reading", *Nature*, Vol. 463, 27 January, pp. 416-18, www.nature.com/news/2010/100127/full/463416a.html.

Luhn, H. (1958), "A business intelligence system", *IBM Journal of Research and Development*, Vol. 2, No. 4, p. 314, http://domino.watson.ibm.com/tchjr/journalindex.nsf/c469af92ea9eceac85256bd50048567c/fc097c29158e395f85256bfa00683d4c!OpenDocument.

Marcus, G. and E. Davis (2014), "Eight (no, nine!) problems with big data", *New York Times*, 6 April, www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html.

Markoff, J. (2012), "Skilled Work, Without the Worker", The iEconomy, Part 6: Artificial Competence, *New York Times*, 18 August.

Mayer-Schönberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray, London.

Mehta, N. (2012), "Knight $440 million loss sealed by rules on cancelling trades", Bloomberg, 14 August, www.bloomberg.com/news/2012-08-14/knight-440-million-loss-sealed-by-new-rules-on-canceling-trades.html.

Merelli, E. and M. Rasetti (2013), "Non locality, topology, formal languages: New global tools to handle large data sets", International Conference on Computational Science, ICCS 2013, *Procedia Computer Science,* No. 18, pp. 90-99, http://dx.doi.org/10.1016/j.procs.2013.05.172.

Merriam-Webster (2014), "Crowdsourcing", *Merriam-Webster.com,* Merriam-Webster, www.merriam-webster.com/dictionary/crowdsourcing, accessed 24 September 2014.

Metz, C. (2010a), "Google's MapReduce patent - No threat to stuffed elephants", *The Register*, 22 February, http://www.theregister.co.uk/2010/02/22/google_mapreduce_patent.

Metz, C. (2010b), "Google blesses Hadoop with MapReduce patent license", *The Register*, 27 April, http://www.theregister.co.uk/2010/04/27/google_licenses_mapreduce_patent_to_hadoop.

MGI (McKinsey Global Institute) (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company, June, www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx, accessed 24 May 2015.

Microsoft (2011), "Microsoft expands data platform with SQL Server 2012, new investments for managing any data, any size, anywhere", *Microsoft News Center*, 12 October, www.microsoft.com/en-us/news/press/2011/oct11/10-12PASS1PR.aspx.

MIT Technology Review (2010), "Twitter mood predicts the stock market", 18 October, www.technologyreview.com/view/421251/twitter-mood-predicts-the-stock-market/.

Muenchen, R. (2012), "The popularity of data analysis software", *r4stats.com*, http://r4stats.com/articles/popularity/, accessed 21 October 2014.

Netflix (2012), "Netflix recommendations: Beyond the 5 stars", Netflix, June, http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5-stars.html, accessed 7 June 2014.

O'Connor, S. (2013), "Amazon unpacked", *FT Magazine*, 8 February, www.ft.com/intl/cms/s/2/ed6a985c-70bd-11e2-85d0-00144feab49a.html#slide0, accessed 24 March 2015.

OECD (2015a), *Digital Economy Outlook 2015*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264232440-en.

OECD (2015b), *Inquiries into Intellectual Property's Economic Impact*, OECD, forthcoming.

OECD (2014a), "Connected televisions: Convergence and emerging business models", *OECD Digital Economy Papers*, No. 231, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jzb36wjqkvg-en.

OECD (2014b), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris.

OECD (2014c), "International cables, gateways, backhaul and international exchange points", *OECD Digital Economy Papers*, No. 232, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jz8m9jf3wkl-en.

OECD (2014d), "Cloud computing: The concept, impacts and the role of government policy", *OECD Digital Economy Papers*, No. 240, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxzf4lcc7f5-en.

OCDE (2013a), "Building blocks for smart networks", OECD Digital Economy Papers, No. 215, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k4dkhvnzv35-en.

OECD (2013b), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'big data'", *OECD Digital Economy Papers*, No. 222, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k47zw3fcp43-en.

OECD (2013c), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, 11 July, C(2013)79, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2012a), "ICT applications for the smart grid: Opportunities and policy implications", *OECD Digital Economy Papers*, No. 190, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9h2q8v9bln-en.

OECD (2012b), "Machine-to-machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

OECD (2011a), Recommendation of the Council on Principles for Internet Policy Making, 13 December, C(2011)154, www.oecd.org/sti/ieconomy/49258588.pdf, accessed 19 May 2015.

OECD (2011b), Terms of Reference for Ensuring the Continued Relevance of the OECD Framework for Privacy and Transborder Flows of Personal Data, DSTI/ICCP/REG(2011)4/FINAL, www.oecd.org/sti/interneteconomy/48975226.pdf.

OECD (2011c), Quality Framework and Guidelines for OECD Statistical Activities, 17 January, www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en, accessed 6 January 2015.

OECD (2009), "Smart sensor networks: Technologies and applications for green growth", *OECD Digital Economy Papers*, No. 167, OECD Publishing, Paris, http://dx.doi.org/10.1787/5kml6x0m5vkh-en.

O'Neil, C. (2013a), "K-nearest neighbors: Dangerously simple", *Mathbabe*, 4 April, http://mathbabe.org/2013/04/04/k-nearest-neighbors-dangerously-simple/.

O'Neil, C. (2013b), "We don't need more complicated models, we need to stop lying with our models", *Mathbabe,* 3 April, http://mathbabe.org/2013/04/03/we-dont-need-more-complicated-models-we-need-to-stop-lying-with-our-models (accessed 7 June 2014).

Paul, R. (2010), "Google's MapReduce patent: What does it mean for Hadoop?", *Arstechnica.com*, 20 January, http://arstechnica.com/information-technology/2010/01/googles-mapreduce-patent-what-does-it-mean-for-hadoop, accessed 10 May 2014.

Pingdom (2013), "The top 100 web hosting countries", 14 March, http://royal.pingdom.com/2013/03/14/web-hosting-countries-2013, accessed 20 April 2014.

Reimsbach-Kounatze, C. (2015), "The proliferation of data and implications for official statistics and statistical agencies: A preliminary analysis", *OECD Digital Economy Working Papers*, http://dx.doi.org/10.1787/5js7t9wqzvg8-en.

Reimsbach-Kounatze, C. (2009), "Towards green ICT strategies: Assessing policies and programmes on ICT and the environment", *OECD Digital Economy Papers*, No. 155, OECD Publishing, Paris, http://dx.doi.org/10.1787/222431651031.

Ried, S. (2011), "Sizing the cloud", *Forrester*, 21 April, http://blogs.forrester.com/stefan_ried/11-04-21-sizing_the_cloud (accessed 21 April 2013).

Russell, S. and P. Norvig (2009), *Artificial Intelligence: A Modern Approach*, 3rd edition, Prentice Hall.

Russom, P. (2011), *Big Data Analytics*, TDWI Best Practices Report, The Data Warehousing Institute, p. 24, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx.

Russom, P. (2007), "BI search and text analytics: New additions to the BI technology stack", TDWI Best Practices Report, The Data Warehousing Institute, Second Quarter 2007.

Scheuer, Mark (2014), "Continuous EEG monitoring in the intensive care unit", *Epilepsia*, Vol. 43 (Suppl. 3), Blackwell Publishing, Inc., pp. 114-27 and 200,

Schubert, L., K. Jefferey and B. Neidecker-Lutz (2010), "The future of cloud computing: Opportunities for European cloud computing beyond 2010", Public Version 1.0, Expert Group Report, European Commission, http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf.

Segal, D. (2011), "The Dirty Little Secrets of Search" *New York Times*, 12 February, www.nytimes.com/2011/02/13/business/13search.html.

Shapiro, C. and H.R. Varian (1999), *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston, Mass.

Shilakes, C. and J. Tylman (1998), *Enterprise Information Portals: Move Over Yahoo! – The Enterprise Information Portal Is on Its Way*, Merrill Lynch, 16 November.

Smith, B. W. (2014), "Automated Vehicles Are Probably Legal in the United States", 1 Tex. A&M L. Rev. 411 (2014), http://ssrn.com/abstract=2303904.

Sol, H. (1987), "Expert systems and artificial intelligence in decision support systems", proceedings of the Second Mini Euroconference, Lunteren, Netherlands, 17-20 November, Springer.

Speier, C., J. Valacich, and I. Vessey (1999), "The influence of task interruption on individual decision making: An information overload perspective". *Decision Sciences* 30, 2, 7 June, pp 337-360, DOI: 10.1111/j.1540-5915.1999.tb01613.x.

Stewart-Smith, H. (2012), "Foxconn chairman compares his workforce to 'animals'", *ZDnet*, 20 January, www.zdnet.com/blog/asia/foxconn-chairman-compares-his-workforce-to-animals/776.

Surowiecki, J. (2011), "A Billion Prices Now", *The New Yorker*, 30 May.

Swanson D.R. (1986), "Undiscovered Public Knowledge", *Library Quarterly*, Vol. 56, pp. 103-18.

The Economist (2014), "Networked manufacturing: The digital future", March, www.economistinsights.com/sites/default/files/EIU%20-%20Siemens%20-%20Networked%20manufacturing%20The%20digital%20future%20WEB.pdf.

The Economist (2012), "High-frequency trading: The fast and the furious", 25 February, www.economist.com/node/21547988.

The Economist (2010), "The tyranny of choice: You choose", 18 December, www.economist.com/node/17723028.

Tresp, V. (2010), *On the Growing Impact of Machine Learning in Industry*, Siemens, www.sics.se/~aho/tor/Volker_Tresp_ToR-101125.pdf.

Ubaldi, B. (2013), "Open government data: Towards empirical analysis of open government data initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

Unwin, A., M. Theus and H. Hofmann (2006), "Graphics of large datasets: Visualising a million", *Statistics and Computing* Series, Springer, Singapore.

Verdone, R. et al. (2008), *Wireless Sensor and Actuator Networks*, Academic Press/Elsevier, London.

VMware (2011), "Business agility and the true economics of cloud computing", business white paper, www.vmware.com/files/pdf/accelerate/VMware_Business_Agility_and_the_True_Economics_of_Cloud_Computing_White_Paper.pdf, accessed 15 February 2015.

Watters, A. (2012), "Embracing the chaos of data", *O'Reilly Radar*, 31 January, http://radar.oreilly.com/2012/01/unstructured-data-chaos.html.

Wilson, J. (2008), *Sensor Technology Handbook*, Newnes/Elsevier, Oxford.

Young, S. (2010), "Obituary of Frederick Jelinek 1932-2010: The pioneer of speech recognition technology", *SLTC Newsletter*, November, www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2r010-11/jelinek/.

Zins, C. (2007), "Conceptual approaches for defining data, information, and knowledge", *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 4, pp. 479-93, 1 February, www.success.co.il/is/zins_definitions_dik.pdf.

Further reading

Anderson, C. (2012), "The man who makes the future: Wired icon Marc Andreessen", *Wired*, 24 April, www.wired.com/2012/04/ff_andreessen/5/, accessed 05 May 2015.

Amazon (2009), "Amazon Elastic MapReduce Developer Guide API", 30 November, http://s3.amazonaws.com/awsdocs/ElasticMapReduce/latest/emr-dg.pdf (accessed 21 April 2014).

Bakhshi, H. and J. Mateos-Garcia (2012), *Rise of the Datavores: How UK Businesses Analyse and Use Online Data*, Nesta, London.

Bullas, J. (2011), "50 fascinating Facebook facts and figures", jeffbullas.com, 28 April, www.jeffbullas.com/2011/04/28/50-fascinating-facebook-facts-and-figures, accessed 14 July 2014.

Esmeijer, J. et al. (2013), "Thriving and surviving in a data-driven society", TNO report, 24 September, http://publications.tno.nl/publication/34610048/xcv74S/TNO-2013-R11427.pdf, accessed 24 September 2013.

Hachman, M. (2012), "Facebook now totals 901 million users, profits slip", *PC Magazine*, 23 April, www.pcmag.com/article2/0,2817,2403410,00.asp.

ISO (International Organization for Standardization) (2009), ISO/IEC Standards 15408-1, 2, 3:2009 – "Information technology – Security techniques – Evaluation criteria for IT security", http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm.

Kan, M. (2013), "Foxconn to speed up 'robot army' deployment", *PCWorld*, 26 June, www.pcworld.com/article/2043026/foxconn-to-speed-up-robot-army-deployment-20000-robots-already-in-its-factories.html.

Kommerskollegium (2014), "No transfer, no trade – The importance of cross-border data transfers for companies based in Sweden", January, www.kommers.se/Documents/dokumentarkiv/publikationer/2014/No_Transfer_No_Trade_webb.pdf, accessed July 2014.

McGuire, T., J. Manyika and M. Chui (2012), "Why big data is the new competitive advantage", *Ivey Business Journal*, July/August, http://iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.VCJ7lPnoQjM, accessed January 2013.

Metha, N. (2012), "Knight $440 million loss sealed by rules on canceling trades", *Bloomberg*, 14 August, www.bloomberg.com/news/2012-08-14/knight-440-million-loss-sealed-by-new-rules-on-canceling-trades.html.

Mivule, K. (2013), "Utilizing noise addition for data privacy: An overview", Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), pp. 65-71, http://arxiv.org/pdf/1309.3958.pdf, accessed 25 March 2015.

Muthukkaruppan, K. (2010), "The underlying technology of messages", *Notes, Facebook*, 15 November, www.facebook.com/notes/facebook-engineering/the-underlying-technology-of-messages/454991608919, accessed 24 March 2015.

Narayanan, A. and V. Shmatikov (2007), "How to break anonymity of the Netflix prize dataset", 22 November, http://arxiv.org/abs/cs/0610105v2.

OECD (2013d), *The OECD Privacy Framework*, OECD Publishing, Paris, www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.

OECD (1997), Recommendation of the Council concerning Guidelines for Cryptography Policy, [C(97)62/FINAL], 27 March, OECD Publishing, Paris, www.oecd.org/internet/ieconomy/guidelinesforcryptographypolicy.htm.

Ohm, P. (2009), "The rise and fall of invasive ISP surveillance", *University of Illinois Law Review,* 1417.

Pfitzmann, A. and M. Hansen (2010), "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management", v0.34, TU Dresden, 10 August, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, accessed 24 March 2015.

Rao, L. (2011), "Index and Khosla lead $11M round in Kaggle, a platform for data modeling competitions", *TechCrunch*, 2 November, http://techcrunch.com/2011/11/02/index-and-khosla-lead-11m-round-in-kaggle-a-platform-for-data-modeling-competitions/.

Warden, P. (2011), "Why you can't really anonymize your data", O'Reilly Strata, 17 May, http://strata.oreilly.com/2011/05/anonymize-data-limits.html.

Chapter 4

Drawing value from data as an infrastructure

This chapter introduces the theoretical foundation for the economic potential of data and discusses key data governance issues that need to be addressed in order to maximise data's potential and reuse across society. It begins by presenting data as an infrastructural resource and a non-rivalrous capital good. It goes on to discuss how data's value depends entirely upon context, with reuse enabling multi-sided markets in which huge returns to scale and scope can lead to positive feedback loops. The often misunderstood notion of "ownership" is discussed, and data quality is seen as multi-faceted and involving seven dimensions. The key aspects of data access, sharing, portability and interoperability are examined and presented as elements of a data governance framework that can help overcome barriers to the reuse of data.

> *I recognised that information was, in many respects, like a public good, and it was this insight that made it clear to me that it was unlikely that the private market would provide efficient resource allocations whenever information was endogenous.* (Stiglitz, 2001)

> *Through ever-expanding commerce, the nation becomes ever-wealthier, and hence trade and commerce routes must be held open to the public, even if contrary to private interest. Instead of worrying that too many people will engage in commerce, we worry that too few will undertake the effort.* (Rose, 1986)

Data have increasingly become an important source of value creation and (data-driven) innovation (DDI). More and more organisations collect, store, and process data today to expand their future production capacities (see Chapter 1 and 2 of this volume), and the productivity improvements are truly dramatic. TomTom, a leading provider of navigation hardware and software, now has more than nine trillion data points collected from its navigation devices and other sources, describing time, location, direction and speed of travel of individual anonymised users, and it now adds six billion measurement points every day.[1] The results of the data analysis are fed back to its navigation devices to inform drivers about current and predicted traffic. This can lead to significant time savings and reduce congestion. Overall, estimates suggest that the global pool of personal geo-locational data has been growing by 20% a year since 2009. By 2020, this data pool could provide USD 500 billion in value worldwide in the form of time and fuel savings, or 380 million tonnes of CO2 emissions saved (MGI, 2011).

As the use of data becomes an increasingly important economic and social phenomenon, economists and policy analysts are trying to capture the phenomenon through existing concepts and theories. Metaphors such as "data is the new currency" (Schwartz, 2000 cited in IPC, 2000; Zax, 2011; Dumbill, 2011; Deloitte, 2013) or, more recently, "data is the new oil" (Kroes, 2012; Rotella, 2012; Arthur, 2013) are often used as rhetorical means to make this emerging phenomenon better understandable to policy and decision makers. Although at first helpful to highlight the (new) economic value of data, these metaphors often fall short and are sometimes even misleading, and therefore should be used with caution (see for example Thorp, 2012; Bracy, 2013; and Glanz, 2013). For example, data are not a rivalrous good, nor are they a primary resource – such as oil, which is depleted once extracted, transformed and burned during production processes. In contrast to oil, the use of data does not exhaust the supply of data and (therefore) *in principle* its potential to meet the demands of others. All these metaphors however reflect an urgent need for a concept through which to better understand and analyse the economics of data, ideally building on familiar concepts, so as to develop better policies and strategies for data's governance.

This chapter responds to that need from a public policy perspective. It provides a framework that can guide policy makers in identifying when data warrant their attention. Not all data are of great value-added from a public policy perspective, at least at first sight: an example here would be data generated when posting on social networks such as Facebook. There are moreover controversies about the use of (e.g.) personal data. However, if the agglomeration and sharing of any data across society can respond to specific societal needs, then that data may merit policy makers' attention. The chapter begins with an analysis of the fundamental economic properties that account for data's potential as a driver of value creation and economic growth and development. These properties include: i) the (non)rivalrous nature of their consumption, ii) their (non)excludability, and iii) the economics of scale and scope in the creation and use of data. These properties lead to the conclusion that data are an infrastructural resource. Building on a rich literature base dealing with the economics of infrastructures, especially the work of Frischmann (2012), the chapter then analyses major supply- and demand-side issues that emerge from data *as* an infrastructure. Special attention is given to potential spillovers (positive externalities) that provide the major theoretical link to total factor productivity growth as highlighted by a number of scholars[2] (among them Corrado et al. [2012]) and the implications in managing data as (knowledge) commons.

4.1. Data as infrastructural resource

The economic properties of data suggest that data may be considered as an infrastructure or infrastructural resource. This may sound counterintuitive, since traditionally infrastructures typically refer to large-scale physical facilities provided for public consumption; the classic examples are transportation systems, including highway and railway systems; communication systems, including telephone and broadband networks; and basic services and facilities such as buildings and sewage and water systems (Frischmann, 2012). However, as for example recognised by the US National Research Council (NRC, 1987), the notion of infrastructure also refers to non-physical facilities, such as education systems and governance systems (including for example the court system). Frischmann (2012) highlights that "the NRC recognised three conceptual needs … first, the need to look beyond physical facilities; second, the need to evaluate infrastructure from a systems perspective; and third, the need to acknowledge and more fully consider the complex dynamics of societal demand". According to Frischmann, the broader concept of infrastructures strongly suggests that they be regarded from a functional perspective rather than from a purely physical or organisational perspective.

As defined by Merriam-Webster, infrastructures are "the basic equipment and structures … that are needed for a country, region, or organisation to function properly". According to Frischmann (2012), they provide the "underlying foundation or basic framework (as of a system or organisation)". That author goes on to state (2012) that infrastructure resources are "shared means to many ends", which satisfy the following three criteria:

1. the resource may be consumed in a non-rivalrous fashion for some appreciable range of demand (i.e. the non-rivalrous criterion)
2. social demand for the resource is driven primarily by downstream productive activities that require the resource as an input (i.e. the capital good criterion)
3. the resource may be used as an input into a wide range of goods and services, which may include private goods, public goods, and social goods (i.e. the general-purpose criterion).

As discussed in the following sections, most (though not all) data are indeed "shared means to many ends" and satisfy Frischmann's three criteria. Therefore, data can in principle be considered an infrastructural resource.

Data as a non-rivalrous good

(Non)rivalry of consumption describes the degree to which the consumption of a resource affects (or does not affect) the potential of the resource to meet the demands of others. It thus reflects the marginal cost of allowing an additional consumer of the good. A purely rivalrous good such as oil can only be consumed once. A non-rivalrous good such as data, in contrast, can be consumed in principal an unlimited number of times. But if this property is, as noted above, the source of significant spillovers that provide the major theoretical link to total factor productivity growth, it also raises questions about how best to allocate data as a resource.

While it is widely accepted that social welfare is maximised when a rivalrous good is consumed by the person who values it the most, and that the market mechanism is generally the most efficient means for rationing such goods and for allocating resources needed to produce such goods, this is not always true for non-rivalrous goods

(Frischmann, 2012). The situation is more complex, since non-rivalrous goods come with an additional degree of freedom with respect to resource management. As Frischmann (2012) highlights, social welfare is maximised not when the good is consumed solely by the person who values it the most, but when everyone who values it consumes it. Maximising access to the non-rivalrous good will in theory maximise social welfare, as every additional private benefit comes at no additional cost.

Data as a capital good

Data are often described as "the new oil". However, besides the non-rivalrous nature of data, there is another drawback with such an analogy: data are neither a consumption good such as an apple, nor an intermediate good such as oil. In most cases, data can be classified as a capital good.

Consumption goods are consumed to generate direct benefits to the consumer or firm. The United Nations System of National Accounts (SNA) defines a consumption good or service as "one that is used (without further transformation in production) by households, NPISHs [non-profit institutions serving households] or government units for the direct satisfaction of individual needs or wants or the collective needs of members of the community" (UN, 2008). In contrast, intermediate goods and capital goods are used as inputs to produce other goods. They are means rather than ends, and their demand is driven by the demand for the derived outputs. They are thus factors of production (see Saviz, 2011; Jones, 2012).

Intermediate consumption is defined by the SNA (UN, 2008) as "consist[ing] of the value of the goods and services consumed as inputs by a process of production, excluding fixed assets whose consumption is recorded as consumption of fixed capital". Capital goods, according to the OECD, are "goods, other than material inputs and fuel, used for the production of other goods and/or services".[3] Intermediate goods such as raw materials (e.g. oil) are used up, exhausted, or otherwise transformed when used as input to produce other goods; capital goods are not. Furthermore, capital goods "must have been produced as outputs from processes of production", which explains why "natural assets such as land, mineral or other deposits, coal, oil, or natural gas, or contracts, leases and licences" are not considered capital goods (UN, 2008).[4]

Data can sometimes be consumed to directly satisfy consumer demand. This is the case for example with an OECD statistic, which will inform the reader about a socio-economic fact. However, in most cases data are not a consumption good but instead are used as an input for goods or services; this is especially true of large volumes of data (i.e. "big data"), which are means rather than ends in themselves. In other words, demand for big data is not for the data itself, but for the benefits that their use promises to bring. In that sense, even pure data products such as infographics (i.e. graphic visual representations of data, information, or knowledge) are the outputs of algorithms applied to data – in the case of infographics, visualisation algorithms.

Data are also not an intermediate good, as they are not exhausted when used given their non-rivalrous nature. This does not mean that data cannot be discarded after they have been used. In many cases, they are used just once. However, while the cost of storing data in the past discouraged keeping data that were no longer, or unlikely to be, needed, storage costs today have decreased to the point where data can generally be kept for long periods of time, if not indefinitely. This has increased data's capacity to be used as a capital good and production factor.

Furthermore, being a capital good does not mean that data do not depreciate like most capital goods, whose value declines "as a result of physical deterioration, normal obsolescence or normal accidental damage" (UN, 2008). In the case of data, depreciation is more complex because it is context dependent, as further described below. Data have no intrinsic value as the value depends on the context of its use. A number of factors presented in more detail in the following sections can affect that value, in particular i) the *accuracy* and ii) the *timeliness* of data. The more relevant and accurate data are for the particular context in which they are used, the more useful and thus valuable data will be (see Oppenheim, Stenson and Wilson, 2004, cited in Engelsman, 2009). This implies, however, that the value of data can perish over time depending on how they are used (see Moody and Walsh, 1999, cited in Engelsman, 2009). Data can especially depreciate in value when they begin to lose their relevance for a particular intended use. There is thus a temporal premium that is motivated by the "real-time" supply of data, for example in the financial sector.

The capital good nature of data has major implications for economic growth. As data are a non-rival capital, they can in theory be used (simultaneously) by multiple users for multiple purposes as an input to produce an unlimited number of goods and services. In practical terms this link to total factor productivity growth finds its application in data-enabled multi-sided markets, i.e. economic platforms in which distinct user groups generate benefits (externalities or spillovers) to other groups.

Data as general-purpose input

As Frischmann explains, "infrastructure resources enable many systems (markets and nonmarkets) to function and satisfy demand derived from many different types of users". They are not inputs that have been optimised for a special limited purpose, but "they provide basic, multipurpose functionality" (Frischmann, 2012). In particular, infrastructures make possible a wide range of private, public and social goods, which users are free to produce according to their capabilities.

How data are used will typically depend on the initial purpose for which they have been collected. For example, at the outset agricultural data will primarily be used for agricultural goods and services. However, in theory there are no limits with regard to the purposes for which data can be used, and many of the benefits stemming from their reuse are based on the fact that data created in one domain can provide further insights when applied in another domain. A clear illustration is provided by open public sector data, where data sets used originally for administrative purposes are reused by entrepreneurs to create services unforeseen when the data were originally created. Likewise, researchers in the areas of health care and Alzheimer's disease are considering reusing retail and social network data to study the impact of behavioural and nutritional patterns on the evolution of the disease.

The general-purpose nature of infrastructure comes with a key policy implication. The production of (ex-ante unforeseeable) public and social goods via the infrastructure could lead to the market failure of insufficient provision of the infrastructure, which would call for government intervention in some cases. As Frischmann explains, "[U]sers' willingness to pay [for the infrastructure] reflects private demand – the value that they expect to realise – and does not take into account value that others might realise as a result of their use" (social value). That "social value may be substantial but extremely difficult to measure", thus leading to a "demand-manifestation problem" which in turn may lead to an undersupply of the infrastructure and a "prioritisation of access and use of

the infrastructure for a narrower range of uses than would be socially optimal" (Frischmann, 2012). As a consequence, there can be significant (social) opportunity costs in limiting access to infrastructures. In other words: open (closed) access enables (restricts) user opportunities and degrees of freedom in the downstream production of private, public and social goods, many of which by their nature have significant spillover effects. In particular, in environments characterised by high uncertainty, complexity and dynamic changes, open access can be an optimal (private and social) strategy for maximising the benefits of an infrastructure.

This means that data markets may not be able to fully serve social demand for data where such a demand manifestation problem would occur. Although no literature is known to have discussed the data demand manifestation problem, there are plausible reasons to believe that such a problem may occur in the data ecosystem, for instance, when data is used to increase transparency in government (see Chapter 10 of this volume). In addition, the context dependency of data and information presented below and the highly uncertain, complex, and dynamic environment in which some data are used (e.g. research) make it almost impossible to fully evaluate ex ante the potential of data, and would exacerbate a demand manifestation problem.

The latter point calls for managing data based on non-discriminatory access regimes, for instance as commons or through open access regimes. Frischmann (2012) points to the following reasons; the first two are in fact closely associated with the concept of *open innovation* (see Box 4.1), as discussed in *The OECD Innovation Strategy* (2010) and the OECD (2013a) project on "Knowledge Networks and Markets":

- *Facilitating joint production or co-operation with suppliers, customers or even competitors* is not a new phenomenon. Joint research ventures or patent pools are well known examples, where firms share resources under non-discriminatory access regimes. This is "because independent research efforts are inhibited by complexity, expense, strategic concerns, transaction costs, or other impediments" (Frischmann, 2012). Sharing agreements are very often an important part of these collaboration efforts. In the particular case of data, access does not need to be open to the public, but it may be limited to the partners who share their data as commons to "overcome collective action problems, sometimes mere co-ordination problems and sometimes more difficult prisoner's dilemma problems"[5] (Frischmann, 2012).
- *Supporting and encouraging value-creating activities by users (user-/consumer-driven innovation)* can be enabled thanks to open access. Open access is an optimal strategy for organisations "when they recognise that users may be best positioned to create value" (Frischmann, 2012). In its weakest form, where users are granted access only to their own personal data, consumers are given "better visibility into their own consumption, often revealing information that can lead to changes in behavior" (MGI, 2013). In its most extreme form, where access is granted to the public, users (including consumers and citizens) are empowered to "provide input to improve the quality of goods and services" (MGI, 2013). This includes improving public services as well as the quality of data.[6]
- *Maximising the option value[7] of the organisation's infrastructural resource when there is high uncertainty regarding sources of future market value.* In contrast with the case described above, where organisations know that users are best placed to create future value, here organisations "are uncertain about the future sources of the value … what unforeseen uses may emerge, what people will want,

how much people will be willing to pay, what complementary goods and services may arise in the future, and so on" (Frischmann, 2012). They adopt open access strategies, taking "advantage of the increased value of experimentation by users, the increased range of potential value-creating services, market selection of the best services that eventually emerge, and learning over time about user preferences and possible paths for continued development". The advantage for the organisation is that it "maintains flexibility and avoids premature optimisation or lock-in to a particular development path or narrow range of paths" (Frischmann, 2012).

- *(Cross-)subsidising the production of public and social goods* requires picking winners (users or applications) by assessing (social) demand for such goods based on the (social) value they create (Frischmann, 2012). Governments can support the production of public goods i) by directly producing these goods, or ii) by supporting private firms' production of public and social goods through (e.g.) research grants, procurement programmes, contracted research and tax incentives. All these strategies raise a number of issues, including difficulties in picking winners and losers, and the fact that resources are limited. Open access regimes can be a more efficient and politically attractive "indirect intervention" to support the production of public and social goods. As Frischmann (2012) highlights, "commons management is not a direct subsidy to … users who produce public or social goods, but it effectively creates cross-subsidies and eliminates the need to rely on either the market or the government to 'pick winners' – that is, to prioritise or rank … users worthy of access and support".

Box 4.1. **Illustrations of "openness"**

Open innovation – This term refers to the "use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation". That includes proprietary-based business models that make active use of licensing, collaborations, joint ventures, etc. Here, "open" is understood to denote the arm's-length flow of innovation knowledge across the boundaries of individual organisations.

Open source – This term is now applied to designate innovations, often jointly developed by different contributors, available royalty-free to anyone and without significant restrictions on how they are to be used. A possible restriction is that derivative work also has to be provided on a same basis.

Open science – This term is often used to describe a movement that promotes greater transparency in the scientific methodology used and data collected; advocates the public availability and reusability of data, tools and materials; and argues for broadly communicating research (particularly when publicly funded) and its results.

Open access – This term refers to the possibility of accessing scientific literature and data "digital, online, free of charge, and free of most copyright and licensing restrictions". This term is also increasingly applied to data provided by profit-driven operators, who develop business models that enable them to obtain a source of revenue bundled alongside information provided on a free and open basis.

Open knowledge – This term coined by the Open Knowledge Foundation refers to any content, information or data that people are free to use, reuse and redistribute, without any legal, technological or social restriction.

Source: OECD (2013a).

4.2. The economics of data

Data increasing returns to scale and scope and network effects

Returns to scale are concerned with changes in the level of output as a result of changes in the amount of factor inputs used. Increasing returns to scale are realised when for example the doubling of the amount of all factors of production results in more than double the output. Returns to scope are conceptually similar to returns to scale, except that it is not the size or the scale of the factor inputs that leads to over-proportionate outputs, but the diversity of the input. In contrast, economies of scale are the cost advantages that organisations obtain thanks to the size of their outputs or the scale of their operation. As the size and scale increases, the cost per unit of output (average cost) decreases. Economies of scope are conceptually similar to economies of scale, except that – once again – it is not the size or the scale of the outputs that leads to over-proportionate reduction in the average cost (cost per unit), but the diversity of the product.

Networks effects, which often referred to as demand-side economies of scale, refer to the fact that the utility of a good to a user (on the demand side) depends on the use of that good by other users. An example often given is the fax machine. While a single fax machine has no utility to a single user, a fax machine starts generating benefits as more users decide to purchase a fax machine, as the technology provides a growing opportunity to communicate with an existing network of users. Many data-driven services and platforms, such as social networking sites, are characterised by large network effects where the utility of the services increases over-proportionately with the number of users. This reinforces the increasing returns to scale and scope on the supply side.

The use of data can generate large returns to scale and scope, as data are non-rivalrous capital that can be reused with positive feedback loops that reinforce each effect at the supply and demand sides. At the same time, the accumulation of data also comes with certain costs (e.g. storage) and risks (e.g. privacy violation and digital security risks). Nevertheless, the advantages are clear:

1. *Increasing returns to scale* – The accumulation of data can lead to significant improvements of data-driven services that in turn can attract more users, leading to even more data that can be collected. This "positive feedback makes the strong get stronger and the weak get weaker, leading to extreme outcomes" (Shapiro and Varian, 1999). For example, the more people use services such as Google Search, or recommendation engines such as that provided by Amazon, or navigation systems such as that provided by TomTom, the better the services, as they become more accurate in delivering requested sites and products and providing traffic information, and the more users they will attract.
2. *Increasing returns to scope* – Diversification of services leads to even better insights if data linkage is possible. This is because data linkage enables "super-additive" insights, leading to increasing returns to scope. Linking data is a means to contextualise data and is thus a source for insights and value that are greater than the sum of isolated parts (data silos). As Newman (2013) highlights in the case of Google: "It's not just that Google collects data from everyone using its search engine. It also collects data on what they're interested in writing in their Gmail accounts, what they watch on YouTube, where they are located using data from Google Maps, a whole array of other data from use of Google's Android phones, and user information supplied from Google's whole web of online services."[8] This diverse data sets enable better profiling hardly possible otherwise.

These effects are not mutually exclusive and may interact, leading to a multiplication. For instance, consumers that appreciate customised search results and ads by Google's search and webmail platform will spend more time on the platform, which allows Google to gather even more valuable data about consumer behaviour and to further improve services, for (new) consumers as well as advertisers (thus on both sides of the market). These self-reinforcing effects may increase with the number of applications provided on a platform, e.g. bundling email, messaging, video, music and telephony – as increasing returns to scope kick in and even more information becomes available thanks to data linkage. As a result, a company such as Google ends up (together with Facebook) with an almost 60% of market share in the US mobile ad market.

Data as non-rivalrous capital enabling multi-sided markets

The effects presented above need to be considered in the context of multi-sided markets that data enable. Two- or multi-sided markets are "roughly defined as markets in which one or several platforms enable interactions between end users and try to get the two or multiple sides 'on board' by appropriately charging each side" (Rochet and Tirole, 2006). These platforms enable multiple distinct groups of customers not only to interact, but also exchange possible externalities among themselves. In other words, the decisions of each group affect the outcome for the other groups. As a consequence, the prices charged to the members of each group will often reflect the effects of these externalities. If the activities of one side create a positive externality for another side (for example more clicks by users on links sponsored by advertisers), then the prices to that other side can be increased (OECD, 2014).

The reuse of data enables multi-sided markets in which huge returns to scale and scope can lead to positive feedback loops in favour of the business on one side of the market, which in turn reinforces success in the other side(s) of the multi-sided market. Established and emerging service platforms such as Google, Facebook, TomTom and John Deere have developed data- and analytics-enabled multi-sided markets, i.e. economic platforms in which distinct user groups generate benefits (externalities or spillovers) for the other side(s). In this they differ from multi-sided markets such as eBay, Amazon, Microsoft's Xbox platform, and Apple's iTunes store. eBay and Amazon, for example, provide online marketplaces for sellers and buyers, and are multi-sided by virtue of their business model (online market). This is also true of Microsoft's Xbox platform, which is positioned in between consumers and game developers, and Apple's iTunes store, which provides a platform that links consumers to application developers and musicians.

In contrast, TomTom's navigation services are provided to consumers as well as to traffic management providers. The service provided to the traffic management providers builds on the analysis of consumer data. The same applies to Google and Facebook, which provide online services to consumers while (re-)using consumer data to provide marketing services to third parties, and to John Deere, which collects agricultural data from farmers and provides them as a service to large seed companies. Data are at the core of these companies' multi-sidedness as non-rivalrous capital collected and used on one side of the market, e.g. to personalise the service, and reused on the other side(s) as input for a theoretically unlimited number of additional goods and services, such as marketing.

Context dependencies

As OECD (2012) highlighted, assessing the value of data ex ante (before use) is almost impossible, because the information derived is context dependent: data that are of good quality for certain applications can thus be of poor quality for other applications. It therefore comes as no surprise that the OECD (2011) Quality Framework and Guidelines for OECD Statistical Activities defines "data quality" as "fitness for use" in terms of user needs, underlining this context dependency (see section below on data quality and curation).

Furthermore, the information that can be extracted from data is not only a function of the data, but also a function of the (analytic) capacity to link data and to extract insights. This capacity is determined by available (meta-)data, analytic techniques and technologies; however, it is a function of pre-existing knowledge and skills. This means that there are factors beyond the data themselves that determine value:

- *Data linkage* – Information depends on how the underlying data are organised and structured. In other words, the same data sets can lead to different information depending on their structure, including their linkages with other (meta-)data.
- *Data analytic capacities* – The value of data depends on the meaning as extracted or interpreted by the receiver. The same data sets can thus lead to different information depending on the analytic capacities of the "receiver", including their skills and (prior) knowledge, available techniques, and technologies for data analysis.

4.3. Towards a data governance framework for better data access, sharing and interoperability

Given their role as the underlying framework of society, infrastructures have always been the object of public policy debates, and governments have played and continue to play a significant and widely accepted role in ensuring the provision of many infrastructures (Frischmann, 2012). The main rationale for the role of governments is justified by the significant spillovers (positive externalities) that infrastructures generate and which result in large social gains, many of which are incompletely appropriated by the suppliers of the infrastructure (Steinmueller, 1996). Spillovers of this nature provide a major theoretical link to total factor productivity growth, but they also present challenges in measuring the contribution of infrastructures or attributing economic growth to that contribution, as the OECD (2012**)** work on measuring the economic impact of the Internet has demonstrated. As Frischmann (2012) explains: "The externalities are sufficiently difficult to observe or measure quantitatively, much less capture in economic transactions, and the benefits may be diffuse and sufficiently small in magnitude to escape the attention of individual beneficiaries."

The positive externalities are also the reason why "infrastructures generally are managed in an openly accessible manner whereby all members of a community who wish to use the resources may do so on equal and non-discriminatory terms" (Frischmann, 2012). The community may, but does not necessarily include the public at large. Furthermore, this does not mean that access is free, nor that access is unregulated. The important point here is that, as Rose highlights (1986, cited in Frischmann, 2012), the positive externalities in combination with open access can lead to a "comedy of the commons", where greater social value is created with greater use of the infrastructure.

Taking commerce as an example, Rose (1986) explains that open access to roads have enabled commerce to generate not only private value that is easily observed and captured by participants in economic transactions, but also social value that is not easily observed and captured by participants (e.g. value associated with socialisation and cultural exchange). In this case, commerce is a productive downstream use of the road infrastructure that generates private as well as social surplus.

In contrast to Hardin's (1968) "tragedy of the commons", where free riding on common (natural) resources leads to the degradation and the depletion of the resources, the "comedy of the commons" is possible in the case of non-rivalrous resources such as data. It is also the strongest rationale for policy makers to promote access to data, either through "open data" in the public sector, "data commons" such as in science, or through the more restrictive concept of "data portability" to empower consumers.

The following section discusses the key challenges related to data governance. These are common challenges that individuals, businesses and policy makers face in every domain in which data are used, irrespective of the type of the data used.

Open data and data commons

A precondition for creating any economic or social value of data is access. Data are a non-rivalrous good and, as mentioned above, their use does not affect in principle their potential to meet the demands of others. As a result, data have unlimited potential to create value. On the other hand, barriers to data access can inhibit data sharing and hinder collaboration, (open) innovation, and the downstream production of data-based goods and services, many of which have significant spillover effects. As a consequence, there can be significant (social) opportunity costs due to barriers to access.

The term "open data" is increasingly used in many different contexts as a solution to promote better access to data. It may actually refer to different concepts, which share a number of commonalities. Open data for governments, for example, often refers to initiatives such as data.gov (United States), data.gov.uk (United Kingdom), or data.gov.fr (France); these enhance access to public sector information (PSI), including public sector data, as encouraged by the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* (see Chapter 10 of this volume).

The term "open data" in the scientific community refers to open access to scientific data, as promoted for example by the OECD (2004) *Declaration on Access to Research Data from Public Funding* and the OECD (2006b) *Council Recommendation concerning Access to Research Data from Public Funding*.[9] All these OECD instruments highlight openness as the first key principle (see Chapter 7 and 10). Last but not least, "open data" is also often associated with movements such as the open source movement, which became particularly popular in the context of open source software (OSS) such as Linux. According to Wikipedia, "Open source as a development model, promotes a) universal access via free license to a product's design or blueprint, and b) universal redistribution of that design or blueprint, including subsequent improvements to it by anyone."

It is important to note that the concept of open data is not limited to the public sector. UN Global Pulse (2012), for example, introduced the concept of "data philanthropy", whereby the private sector shares data to support more timely and targeted policy action, and to highlight the public interest in shared data. In this context two ideas are debated: i) the "data commons", where some data are shared publicly after adequate anonymisation and aggregation; and ii) the "digital smoke signals", where sensitive data

are analysed by companies but the results are shared with governments. The Open Data Institute (ODI), a not-for-profit organisation based in the United Kingdom, is also promoting the release of open data in the private sectors, including but not limited to finance and health care.

Most definitions for open data point to a number of criteria or "principles". According to the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding*, for example, openness means i) access that should be granted on equal or non-discriminatory terms, and ii) access costs that should not exceed the marginal cost of dissemination. As another example, at a meeting of open data advocates in 2007,[10] participants agreed on "8 Principles of Open Government":

- *Complete* – All public data are made available. Public data are data that are not subject to valid privacy, security or privilege limitations.
- *Primary* – Data are as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
- *Timely* – Data are made available as quickly as necessary to preserve their value.
- *Accessible* – Data are available to the widest range of users for the widest range of purposes.
- *Machine processable* – Data are reasonably structured to allow automated processing.
- *Non-discriminatory* – Data are available to anyone, with no registration requirement.
- *Non-proprietary* – Data are available in a format over which no entity has exclusive control.
- *Licence-free* – Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Other definitions that followed focused on a smaller set of criteria. The Open Data White Paper of the United Kingdom Cabinet Office (2012), for example, highlights three of the principles listed above as criteria for open data: i) "accessible (ideally via the Internet) at no more than the cost of reproduction, without limitations based on user identity or intent", ii) "in a digital, machine readable formation for interoperation with other data", and iii) "free of restriction on use or redistribution in its licensing". A recent report by MGI (2013), which defines open data as "the release of information by government and private institutions and the sharing of private data to enable insights across industries", also based its definition on these three criteria, highlighting however access costs as a fourth criterion. A comprehensive discussion of the principles governing open data can be found in Ubaldi (2013).

Among the criteria listed in the above definitions, non-discriminatory access (or "access on equal terms", as stated in the OECD [2005] Recommendation on Principles and Guidelines for Access to Research Data from Public Funding) is central to open data. Non-discriminatory access is about "terms that do not depend on the users' identity or intended use" (Frischmann, 2012; see also United Kingdom Cabinet Office, 2012). As highlighted above, access independent of identity and intent can be crucial for

maximising the value of data across society, as it keeps the range of opportunities as wide as possible.

All other criteria listed above are factors affecting the level of non-discriminatory access, and thus the degree of openness. Three criteria deserve to be highlighted, as they significantly affect the degree of openness (ordered by their increasing magnitude of influence):

- *Technological design* is a broad concept that includes all technical aspects affecting the (re-)use and distribution of data. These factors were presented in Berners-Lee's (2006b) proposed "5 Star Deployment Scheme for Open Data": 1) "make your stuff available on the Web (whatever format) [under an open licence]"; 2) make it available as structured data (e.g. Excel instead of an image scan of a table); 3) "use non-proprietary formats (e.g., CSV [comma-separated values] instead of Excel)"; 4) "use URIs [uniform resource identifiers] to identify things, so that people can point at your stuff"; 5) "link your data to other data to provide context". In essence, the scheme points to the following key technological factors affecting the degree of data openness: i) data availability (ideally online), ii) machine readability (of structured data), and iii) data linkability. It should be noted that factor (i) is required for factor (ii), which in turn is a requirement for factor (iii).
- *Intellectual property rights (IPRs)* – Data can be subject to legal regimes, copyright as well as other IPRs applicable to databases (Box 4.2.) and trade secrets, which need to be respected as highlighted in the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information*. These rights can in some cases limit or prevent the (re-)use and distribution of open data. Some open data initiatives therefore explicitly state that open data should be free of any IPRs (see the 8 Principles of Open Government above). In other cases, innovative IP regimes are used and even promoted through open data regimes, as long as they do not restrict the rights of users to reuse and sometimes redistribute the data. In 2010, for example, the United Kingdom created the *Open Government Licence*[11] to release public sector information (including data) for free without restricting (re-)use or distribution, with the only requirement being attribution. This new licence scheme was based on the *Creative Commons* (CC) licences, another licence scheme widely used for open data.[12] Another example of open licence schemes used for data is the Open Data Commons Open Database License (ODbL), which is for example used for OpenStreetMap data.[13] (For further discussion on IPRs see OECD [2015], Inquiries into Intellectual Property's Economic Impact, OECD, forthcoming).
- *Pricing* – Although pricing will have less of an impact on the degree of openness than technological design and IPRs, it can nevertheless be one of the most challenging factors, because optimal pricing can be hard to determine. Many governments, for example, wish to engage in cost recovery, partly for budgetary reasons and partly based on the principle that those who benefit should pay. But the calculation of benefits can be problematic due to significant spillover effects through the creation of public and social goods based on open data. Furthermore, as Stiglitz et al. (2000) have argued, if government provision of a data-related service is a valid role, generating revenue from that service is not. Many open data initiatives therefore encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" as stated in the OECD (2005)

Recommendation on Principles and Guidelines for Access to Research Data from Public Funding. The OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* further specifies that "where possible, costs charged to any user should not exceed marginal costs of maintenance and distribution, and in special cases extra costs for example of digitisation". While marginal cost pricing is often considered the best option for the public sector, that option is seen as unattractive for the private sector, for which at least cost recovery is a necessity. This can lead to average cost pricing as an alternative pricing model, or can even require complex revenue models including subscription fees, freemium[14] and voluntary donations, in combination with cross-subsidies.

Box 4.2. **Database protection**

Databases are protected by copyright under certain circumstances, but in some countries – namely in the European Union, Japan and South Korea – they are also protected by a so-called sui generis database right (SGDR) aimed at protecting the investment.

The Berne Convention does not mention databases, but provides protection for collections of literary or artistic works such as encyclopaedias and anthologies that, by reason of the selection and arrangement of their contents, constitute intellectual creations.[1] The plain meaning of that provision seems to exclude from protection collections that do not consist of works, which is to say that collections of data (databases) are not covered by Art. 2(5). It has been argued that collections of data are in fact covered by the general provision of Art. 2(1) as "literary and artistic works".

In any event, currently the protection afforded to databases (as collections of data or other elements) is established – or confirmed – by both Art. 10(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) and the almost identical Art. 5 of the WIPO Copyright Treaty: "Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creation shall be protected as such..."[2]

An additional layer of protection is found in some countries and is afforded to databases regardless of the intellectual creation (i.e. "selection or arrangement") that may or may not be present. What is protected here is the investment in generating the database, i.e. in the obtaining, verification or presentation of the data. This type of right, also known as the sui generis database right mentioned above, is found in the EU Database Directive and the laws of a number of other countries, and will be dealt with below. It should be borne in mind that while the protection afforded to original databases focuses on the arrangement or selection without extending to the content of the database, the SGDR offers protection against the copying of substantial parts of the database – that is to say it extends, at least to some extent, to the data themselves.

1. See Art. 2(5) of the Berne Convention available at www.wipo.int/treaties/en/text.jsp?file_id=283698.

2. See Art. 10(2) of the TRIPS Agreements at www.wipo.int/wipolex/en/other_treaties/text.jsp?file_id=305907.

Source: OECD (2015), *Inquiries into Intellectual Property's Economic Impact?* (forthcoming), Chapter 7, "Legal Aspects of Open Access to Publicly Funded Research".

The three factors presented above (technological design, IPRs and pricing) determine the degree of openness, which can range from *closed* (access only by the data controller) to *open to the public* at its two extremes. In between, access may be restricted to i) individual stakeholders who can affect or are affected by the use of the data, with

access typically being granted on discriminatory bases, and to ii) specific communities (see the OECD 2005 Recommendation on Principles and Guidelines for Access to Research Data from Public Funding), with access being restricted to the "international research community". This leads to a three-level definition of open access, as illustrated in Figure 4.1.

Figure 4.1. **The data common continuum**



Overall, open data can be an optimal (private and social) strategy for maximising the benefits of data, in particular in environments characterised by high uncertainty, complexity and dynamic evolution such as climate change, urban development and health care research. These complex systems are often characterised by complementary effects; non-discriminatory access can be a means of internalising them by encouraging "experimentation and innovation among complementary applications" (Frischmann, 2012).

There are a number of other factors affecting the degree of openness: confidentiality and privacy considerations may be justifications for limiting data access in some cases as well. Furthermore, access problems and issues at the international level can emerge due to differences in culture and legislations. OECD (2013d) discusses the following factors in the particular context of science, but they are valid for other domains as well:

- *Legal and cultural barriers* – Depending upon the perceived sensitivity of the data and/or the legal framework governing data-sharing arrangements, some departmental "gatekeepers" can regulate access conditions tightly.
- *Public concerns* – To date there has been relatively little public engagement to explain the potential of data linkage, or the methods that are used to protect individual confidentiality when such linkages are made.
- *Technical barriers* – While various models for secure data access exist in some countries, the expertise, hardware and software to implement secure access is unevenly distributed among countries.

Finally, the provision of high-quality data can require significant time and up-front investments before the data can be shared. These include the costs related to i) datafication, ii) data collection, iii) data cleaning and iv) data curation. Effective knowledge sharing is, however, not limited to sharing data. In many cases a number of complementary resources may be required, ranging from additional (meta-)data to data models and algorithms for data storage and processing, and even secured IT infrastructures for (shared) data storage, processing, and access. For example, data from the distributed array telescope may create large data sets, which however require additional data on the direction of the telescopes to be interpreted correctly.

Given these significant costs, creators and controllers of data do not necessarily have the incentives to share their data. One reason is that the costs of data sharing are perceived as higher than the expected private benefits of sharing. Also, since data are in principle non-exclusive goods for which the costs of exclusion can be high, there is the possibility that some may "free ride" on others' investments. The argument that follows is that if data are shared, free-riding users can "consume the resources without paying an adequate contribution to investors, who in turn are unable to recoup their investments" (Frischmann, 2012). In science and research the situation poses even more incentive problems, as scientists and researchers traditionally compete to be first to publish scientific results, and may (a third disincentive) not enjoy or even perceive the benefits of disclosing the data they could further use for as yet uncompleted research projects (see Chapter 7 of this volume).

The root of these incentive problems can be summarised as a positive externality issue: data sharing may benefit others more than it benefits the data creator and controller, who cannot privatise these benefits and as a result may not sufficiently invest in data sharing or may even refrain completely. However, the idea that positive externalities and free riding always diminish incentives to invest has been challenged by some:

> *There is a mistaken tendency to believe that any gain or loss in profits corresponds to an equal or proportional gain or loss in investment incentives, but this belief greatly oversimplifies the decision-making process and underlying economics and ignores the relevance of alternative opportunities for investment. The conversion of surplus realised by a free rider into producer surplus may be a wealth transfer with no meaningful impact on producers' investment incentives or it may be otherwise, but there is no theoretical or empirical basis for assuming that such producer gains are systematically incentive-relevant.* (Frischmann, 2012)

Such an assumption therefore cannot be generalised, and needs careful case-by-case scrutiny. Indeed, free riding is sometimes the economic and social rationale for providing access to data. Open data, for example, is motivated by the recognition that users will free ride on the data provided, and in so doing will be able to create a wide range new goods and service that were not anticipated and otherwise would not be produced. In that sense, according to Frischmann, "free riding is pervasive in society and a feature, rather than a bug" (2012).

Data portability and interoperability

Data are rarely harmonised across sectors or organisations, as individual units collect and/or produce their own set of data using different metadata, formats and standards. Even if access to data is provided, the data may not be able to be reused in a different context for new applications. Reusability will typically be limited if data are not machine readable and cannot be reused across IT systems (interoperability). Some data formats that are considered machine readable are therefore based on open standards, such as RDF (Resource Description Framework), XML (eXtensible Markup Language), and more recently JSON (JavaScript Object Notation). Other standards include file formats such as CSV (comma-separated values) and proprietary file formats such as Microsoft Excel. Unresolved interoperability issues are, for example, still high on the e-government agendas of many OECD countries (see Chapter 7 of this volume). For instance, interoperability of data catalogues, or the creation of a pan-European data catalogue, is an important challenge currently faced by EU policy makers.

An important development in the context of data portability and interoperability is the increasing role of consumers in the data-sharing ecosystems. In enabling their personal data to flow across organisations, consumers are playing an important role that derives from their access to their own data under the Individual Participation Principle of the OECD (2013b) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines). Furthermore, the Individual Participation Principle grants individuals the right "to challenge data relating to [them] and, if the challenge is successful to have the data erased, rectified, completed or amended" (subject to regulatory obligations, e.g. to keep billing information, etc.). This is a right they could exert when porting their data from one controller to another.

Government initiatives are promoting data portability and thus contributing to the free flow of data as well. In 2011, a government-backed initiative called midata was launched in the United Kingdom to help individuals access their transaction and consumption data in the energy, finance, telecommunications and retail sectors. Under the programme, businesses are encouraged to provide their customers with their consumption and transaction data in a portable, preferably machine readable format. A similar initiative has been launched in France by Fing (Fondation Internet Nouvelle Génération), which provides a web-based platform, MesInfos,[15] for consumers to access their financial, communication, health, insurance and energy data that are being held by businesses. Both the UK and French platforms are outgrowths of ProjectVRM,[16] a US initiative launched in 2006 that provides a model for Vendor Relationship Management by individual consumers. Last but not least, the right to data portability proposed by the European Commission in the current proposal for reform of its data protection legislation aims at stimulating innovation through more efficient and diversified use of personal data, by allowing users "to give their data to third parties offering different value-added services" (EDPS, 2014).

The initiatives discussed above show promise in terms of helping individuals make informed decisions and increasing trust in the data-intensive services that organisations seek to deliver. But such programmes may also bring significant costs with regard to both developing and maintaining the mechanisms for enhanced data access and complying with relevant regulations (Field Fisher Waterhouse, 2012). The question arises: who should bear these costs?

Data linkage and integration

The value of data is, as stated above, highly context dependent – it increases when the data can be linked with and integrated into other data sets. As data are placed in a larger context, they can reveal additional insights that otherwise were not possible to gain. This is for instance true with linked micro data sets, as the example of the Micro-Data Lab of the OECD Directorate for Science, Technology and Innovation (DSTI) demonstrated, where data on firms' innovation performance (e.g. patent applications) are linked with data on their economic performance (e.g. financial statements). Linked data thus create super-additive value, which is greater than the sum of its parts (i.e. of data silos).

There are various reasons why linking data across different silos may be challenging. Some are obviously related to the legal, cultural and technical barriers to data access and sharing, as highlighted above. Others may be related to skills barriers. As OECD (2013d) highlights: "even though techniques for record linkage are now well developed, and are used by numerous organisations regularly, the capacity with which to carry out successful

linkages may be in short supply". Also, some of the barriers to data linkage are legitimate, since linkage can undermine privacy protective measures such as anonymisation and pseudonymisation, as highlighted in Chapter 5 of this volume.

Data quality and curation

The information that can be extracted from data depends on the quality of the data, and data quality in turn depends on the intended use. "If data [are] accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data" (OECD, 2011). Thus, data quality needs to be viewed as a multi-faceted concept. The OECD (2011) defines the following seven dimensions:

1. Relevance – "is characterised by the degree to which the data [serve] to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts".
2. *Accuracy* – is "the degree to which the data correctly estimate or describe the quantities or characteristics they are designed to measure".
3. *Credibility* – "the credibility of data products refers to the confidence that users place in those products based simply on their image of the data producer, i.e. the brand image. Confidence by users is built over time. One important aspect is trust in the objectivity of the data".
4. *Timeliness* – "reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon. … Real-time data [are] data with a minimal timeliness".
5. *Accessibility* – "reflects how readily the data can be located and accessed", as discussed in the previous section on data access and sharing.
6. *Interpretability* – "reflects the ease with which the user may understand and properly use and analyse the data". The availability of metadata plays an important role here, as they provide for example "the definitions of concepts, target populations, variables and terminology, underlying the data, and information describing the limitations of the data, if any".
7. *Coherence* – "reflects the degree to which they are logically connected and mutually consistent. Coherence implies that the same term should not be used without explanation for different concepts or data items; that different terms should not be used without explanation for the same concept or data item; and that variations in methodology that might affect data values should not be made without explanation. Coherence in its loosest sense implies the data are 'at least reconcilable'".

The OECD Privacy Guidelines also provides a number of criteria for data quality in the context of privacy protection. The Recommendation states that "personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date". So the data quality dimensions would have to include completeness as an eighth dimension according to the OECD Privacy Guidelines. Furthermore, the cost efficiency with which data are collected could also be considered as a measure for data quality. "Whilst the OECD does not regard

cost-efficiency as a dimension of quality, it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions" (OECD, 2011).

Data curation embodies those data management activities needed to assure long-term data quality across the data life cycle. Data curation thus includes activities affecting the eight dimensions of data quality presented above. As OECD (2013c) highlights, however, "these particular activities [...] are often beyond the scope and timeframe of original [...] projects" for which the data were initially collected and used. This can lead to disincentives for data curation and put at risk long-term access and reuse of data. In science and research, where the long-term quality of data is essential, data curation is seen as a key part of the provision of research infrastructure (OECD, 2013c).

Data ownership and control

Data ownership is a concept that is often misunderstood and/or misused. With businesses, for example, data ownership is often used to assign responsibility and accountability for specific databases (the "data owners"). In this context, ownership is perceived as a means of assuring data quality and curation, as well as data protection and security along the complete data life cycle. However, ownership is assigned without IPRs being granted to the "data owner" (Scofield, 1998; Chisholm, 2011). Scofield (1998) therefore suggests replacing the term "ownership" with "stewardship", as this better captures the responsibility that organisations are actually looking to promote with the ownership concept.

Granting private property rights is often suggested as a solution to the incentive problems related to free riding. The concept of ownership typically means "to have legal title and full property rights to something" (Chisholm, 2011). Data are an intangible asset; like other information-related goods, they can be reproduced and transferred at almost zero marginal costs. So in contrast to the concept of ownership of physical goods, where the owner typically has exclusive rights and control over the good – including for instance the freedom to destroy the good – this is not the case for intangibles such as data. For these types of goods, IPRs are typically suggested as the legal means to establish clear ownership. In the case of data in particular, legal regimes such as copyright as well as other IPRs applicable to databases and trade secrets can be used (see Box 4.2). Furthermore, technologies such as cryptography have dramatically reduced the costs of exclusion, and thus are often used as a means to protect data (see Chapter 5 of this volume).

However, in contrast to other intangibles, data typically involve complex assignments of different rights across different data stakeholders, requiring of some stakeholders "the ability to access, create, modify, package, derive benefit from, sell or remove data, but also the right to assign these access privileges to others" (Loshin, 2002). So in many cases, no single data stakeholder will have exclusive rights. Different stakeholders will typically have different powers depending on their role. As Trotter (2012) highlights in the case of health patient data, all stakeholders (including patient, doctor and programmer) "have a unique set of privileges that do not line up exactly with any traditional notion of 'ownership'. Ironically, it is neither the patient nor the [doctor] who is closest to 'owning' the data. The programmer has the most complete access and the only role with the ability to avoid rules that are enforced automatically by electronic health record (EHR) software". Loshin (2002) identifies the following data stakeholders that could claim data ownership:

- *creator* – the party that creates or generates data
- *consumer* – the party that uses the data
- *compiler* – the party that selects and compiles information from different information sources
- *enterprise* – all data that entering the enterprise or created within the enterprise is completely owned by the enterprise
- *funder* – the user that commissions the data creation and therefore claims ownership
- *decoder* – in environments where information is 'locked' inside particular encoded formats, the party that can unlock the information becomes an owner of that information"
- *packager* – the party that collects information for a particular use and adds value through formatting the information for a particular market or set of consumers
- *reader as owner* – the value of any data that can be read is subsumed by the reader and, therefore, the reader gains value through adding that information to an information repository
- *subject as owner* – the subject of the data claims ownership of that data, mostly in reaction to another party claiming ownership of the same data
- *purchaser/licenser as owner* – the individual or organisation that buys or licenses data may stake a claim to ownership.

In cases where the data are considered "personal data" the situation is even more complex, since certain rights of the data subject cannot be waived. For example, the Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* recommends that individuals have "the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him within a reasonable time; […] c) to be given reasons if a request made under sub-paragraphs (a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him …". The rights of the data subject limit any possibility for exclusive right on the storage and use of the data.

There are also economic reasons why granting private property rights may not be the optimal solution in the case of data. As highlighted above, social welfare is maximised when a rivalrous good is consumed by the person who values it the most, while social welfare through the consumption of non-rivalrous goods is maximised when the good is consumed by everyone who values it. This additional degree of freedom suggests that other institutions such as commons and "data citations" (see Chapter 7 of this volume) may be more effective in maximising welfare while still providing sufficient incentive for the production and release of data. Furthermore, the free riding story can be "translated in game-theoretic terms into a prisoners' dilemma, another good story, although one that does not necessarily point to private property as a solution to the cooperation dilemma" (Frischmann, 2012).

Overall, "the concept [of ownership] doesn't map well to the people and organisations that have relationships with that data" (Trotter, 2012). Data ownership can be a poor starting point for data governance, and can even be misleading. As Croll (2011) points out: "The important question isn't who owns the data. Ultimately, we all do. A better

question is who owns the means of analysis? Because that's how […] you get the right information in the right place. The digital divide isn't about who owns data – it's about who can put that data to work".

Data value and pricing

The discussion has underlined that data have no intrinsic value; their value depends on the context of their use. In fact, information – more than any other good – is an experience good, i.e. a good that consumers must experience in order to value. "Virtually any new product is an experience good"; however, "information is an experience good every time it's consumed" (Shapiro and Varian, 1999). Data pricing schemes can thus be complex. In particular, the context dependency of data challenges the applicability of market-based pricing: that pricing assumes that markets can converge towards a price at which demand and offer meet, and such is not always the case.

As the OECD (2012) study "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value" showed, the monetary valuation of the same data set can diverge significantly among market participants. For example, while economic experiments and surveys in the United States indicate that individuals are willing to reveal their social security numbers for USD 240 on average, the same data sets can be obtained for less than USD 10 from US data brokers such as Pallorium and LexisNexis. Data pricing schemes based on cost structure seem to be a more common approach. As noted above, the OECD (2005) *Recommendation on Principles and Guidelines for Access to Research Data from Public Funding* – and the OECD (2008) *Council Recommendation on Enhanced Access and More Effective Use of Public Sector Information* – both encourage the provision of data "at the lowest possible cost, preferably at no more than the marginal cost" which can include the cost for "maintenance and distribution, and in special cases extra costs for example of digitisation".

4.4. Key findings and policy conclusions

Data are an infrastructural resource – a capital good that cannot be depleted and that can be used for a theoretically unlimited range of purposes. In particular, data enable multi-sided markets – which, combined with increasing returns to scale and scope – provide businesses with significant growth opportunities (see Chapter 4 of this volume). There are, however, data demand manifestation problems, which may lead to under-provision of data or the prioritisation of access and use for a narrower range of uses than would be socially optimal.

This calls for managing data based on non-discriminatory access regimes, including commons or open access regimes, because:

1. these regimes facilitate joint production or co-operation with suppliers, customers or even competitors
2. they support and encourage value-creating activities by users
3. they maximise the option value of data and data-related products when there is high uncertainty regarding sources of future market value
4. they are (cross-)subsidising the production of public and social goods, which otherwise would require governments or businesses to pick winners (users or applications) by assessing the right (social) demand for such goods based on the (social) value they create.

The provision of high-quality data can require significant up-front investments. These costs can sometimes exceed the private benefits expected from data sharing, and thus present a barrier to data sharing. The possibility of "free riding" on others' investments is sometimes seen as a source of additional incentive problems, although there are many cases where free riding had no significant disincentive effects on producing or sharing data (e.g. open data).

"Ownership" is a questionable appellation when it comes to data. In contrast to other intangibles, data typically involve complex assignments of different rights across different data stakeholders. Those different stakeholders will typically have different power over the data, depending on their role. In cases where the data are considered "personal data", the concept of data ownership by the party that collects personal data is even less practical since privacy regimes grant certain explicit control rights to the data subject, as for example specified by the Individual Participation Principle of the OECD (2013c) *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*.

Lack of data portability and interoperability are among the most challenging barriers to data reuse. This is particularly the case where data are not provided in a machine readable format and thus cannot be reused across IT systems. Individuals (consumers) play an important role in promoting the free flow of their personal data across organisations. Government and private sector initiatives such as midata (United Kingdom), MesInfos (France), and the proposed reform of EU data protection legislation are promoting data portability – and thus promoting the free flow of data across organisations – as a means of empowering individuals and consumers and strengthening their participation in DDI processes.

Even within organisations, especially large ones, data silos are perceived as a barrier to intra-organisational data sharing. According to a survey by the Economist Intelligence Unit (2012a), almost 60% of companies stated that "organisational silos" are the biggest impediment to using "big data" for effective decision making. Executives in large firms (with annual revenues exceeding USD 10 billion) are more likely to cite data silos as a problem (72%) than those in smaller firms (with revenues less than USD 500 million, 43%).

Better data governance regimes are needed to overcome barriers to data access, sharing and interoperability (subject to legitimate restrictions, such as privacy). These barriers are often faced by individuals, businesses and policy makers alike across sectors. Data governance regimes can have an impact on the incentives to share and the possibility of data to be used in interoperable ways. The elements to be considered for an effective data governance regime include:

- data access and reuse
- data portability and interoperability
- data linkage and integration
- data quality and curation
- data "ownership" and control
- data value and pricing.

Coherent guidelines are needed to promote better data governance across the economy. Many of the barriers to data access and reuse, for example, are common across

domains, including science and research (Chapter 7 of this volume), health care (Chapter 8) and smart cities (Chapter 9), and the public sector (see Chapter 10). Existing frameworks that promote better access to data, some of which are sector specific, may need to be reviewed and eventually consolidated to foster coherence among public policies related to data access, linkage and reuse. This would also include the OECD Council Recommendations promoting better access to data, including in particular the OECD (2008) *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* of 30 April 2008, and the OECD (2006b) *Recommendation of the Council concerning Access to Research Data from Public Funding* of 14 December 2006, both of which are currently under review.

Notes

1 See www.tomtom.com/en_gb/licensing/products/traffic/historical-traffic/custom-travel-times.

2 See Klenow and Rodriguez-Clare (2005) for an excellent review of the most relevant theoretical models of technological spillovers and economic growth.

3 See OECD, *Main Economic Indicators*, "Sources and Definitions", http://stats.oecd.org/mei/default.asp?lang=e&subject=1, accessed 9 April 2015.

4 The System of National Accounts uses the term "fixed capital" (in contrast to circulating capital, such as raw materials) to refer to capital goods.

5 The prisoner's dilemma is a central part of game theory. It describes a game with two players ("prisoners") that have the opportunity to collaborate to achieve a high payout, or to betray each other for a lower payout. Both players make their choice wihout knowing the choice of the other player, and in case of no collaboration, it is the player who betrays that profits strongly. The prisoner's dilemma is therefore used to illustrate why "rational" individuals might not cooperate, even if collaboration is in their best interests.

6 Altogether, over 50% of the total potential value of open data (more than USD 3 trillion annually) is estimated to be generated from consumer and customer surplus (MGI, 2013). The total value of open data must exceed by far the benefits highlighted in MGI (2013), which attributes the largest share of the total benefits of open data to better benchmarking, "an exercise that exposes variability and also promotes transparency within organizations" (MGI, 2013). Better benchmarking would enable "fostering competitiveness by making more information available and creating opportunities to better match supply and demand" as well as "enhancing the accountability of institutions such as governments and businesses [to] raise the quality of decision [making] by giving citizens and consumers more tools to scrutinize business and government" (MGI, 2013).

7 "Costs and benefits are rarely known with certainty, but uncertainty can be reduced by gathering information. Any decision made now and which commits resources or generates costs that cannot subsequently be recovered or reversed, is an irreversible decision. In this context of uncertainty and irreversibility it may pay to delay making a decision to commit resources. The value of the information gained from that delay is the option value or quasi-option value." (OECD, 2006a)

8 The "super-additive" nature of linked data is of course not without its challenges as well. In particular, linked data sets can undermine confidentiality and privacy protection measures such as anonymisation and pseudonymisation.

9 See also OECD (2005) *Principles and Guidelines for Access to Research Data from Public Funding*, www.oecd.org/sti/sci-tech/38500813.pdf, accessed 12 June 2014.

10 The meeting was organised by Tim O'Reilly of O'Reilly Media and Carl Malamud of Public. Resource.Org. See https://public.resource.org/8_principles.html, accessed 7 November 2013.

11 See www.nationalarchives.gov.uk/doc/open-government-licence/version/2/.

12 See data.australia.gov.au, data.gv.at, and *Google Ngram Viewer,* http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

13 See www.openstreetmap.org/copyright.

14 This business model offers free service to customers, and a premium level of the service is available for a fee (see for example Dropbox).

15 See: http://fing.org/?-MesInfos-les-donnees-personnelles-&lang=fr.

16 See: http://cyber.law.harvard.edu/projectvrm/Main_Page.

References

Arthur, C. (2013), "'Data is the new oil': Tech giants may be huge, but nothing matches big data", *Raw Story*, 24 August, www.rawstory.com/rs/2013/08/24/data-is-the-new-oil-tech-giants-may-be-huge-but-nothing-matches-big-data/, accessed 6 April 2015.

Berners-Lee, T. (2006a), "Isn't it semantic?", www.bcs.org/content/conWebDoc/3337, accessed 4 May 2015.

Berners-Lee, T. (2006b), "Linked data", www.w3.org/DesignIssues/LinkedData, accessed 6 April 2015.

Bracy, J. (2013), "Changing the conversation: Why thinking 'Data Is the New Oil' may not be such a good thing", *Privacy Perspectives*, International Association of Privacy Professionals (IAPP), 19 July, https://privacyassociation.org/news/a/changing-the-conversation-why-thinking-data-is-the-new-oil-may-not-be-such/, accessed 6 April 2015..

Chisholm, M. (2011), "What is data ownership?", www.b-eye-network.com/view/15697, accessed 5 November 2014.

Corrado, C., C. Hulten and D. Sichel (2009), "Intangible Capital and U.S. Economic Growth", *Review of Income and Wealth*, Series 55, No. 3, September, www.conference-board.org/pdf_free/IntangibleCapital_USEconomy.pdf.

Croll, A. (2011), "Who owns your data", 11 January, http://news.yahoo.com/owns-data-20110112-030058-029.html, accessed 5 November 2014.

Deloitte (2013), "Data as the new currency: Government's role in facilitating the exchange", *Deloitte Review*, Issue 13, 24 July, http://cdn.dupress.com/wp-content/uploads/2013/07/DR13_data_as_the_new_currency2.pdf.

Dumbill, E. (2011), "Data is a currency: The trade in data is only in its infancy", *O'Reilly Radar*, http://radar.oreilly.com/2011/02/data-is-a-currency.html, accessed 6 April 2015.

EDPS (European Data Protection Supervisor) (2014), "Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the Digital Economy", March, https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2014/14-03-26_competitition_law_big_data_EN.pdf, accessed 6 April 2015.

Engelsman, W. (2009), "Information assets and their value", University of Twente, http://referaat.cs.utwente.nl/conference/6/paper/6807/information-assets-and-their-value.pdf, accessed 6 April 2015.

Field Fisher Waterhouse (2012), "Will access to midata work?", http://privacylawblog.ffw.com/2012/will-access-to-midata-work, accessed 6 April 2015.

Frischmann, B.M. (2012), *Infrastructure: The Social Value of Shared Resources*, Oxford University Press.

Glanz, J. (2013), "Is big data an economic big dud?", *New York Times*, 17 August, www.nytimes.com/2013/08/18/sunday-review/is-big-data-an-economic-big-dud.html, accessed 6 April 2015.

Hardin, G. (1968), "The tragedy of the commons", *Science* (American Association for the Advancement of Science, AAAS) Vol. 162, No. 3859, pp. 1243-48, www.sciencemag.org/content/162/3859/1243.full.pdf.

IPC (Information and Privacy Commissioner of Ontario) (2000), "Should the OECD Guidelines apply to personal data online?", Report to the 22nd International Conference of Data Protection Commissioners, Venice, Italy, September, www.ipc.on.ca/images/resources/up-oecd.pdf, accessed 6 April 2015.

Jones, S. (2012), "Why 'Big Data' is the fourth factor of production", *Financial Times*, 27 December, www.ft.com/intl/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html, accessed 26 January 2013.

Klenow, P. and A. Rodriguez-Clare (2005), "Externalities and Growth", in A. Philippe and S. Durlauf ed, *The Handbook of Economic Growth*, Elsevier, Amsterdam.

Kroes, N. (2012), "Digital agenda and open data: From crisis of trust to open governing", European Commission, SPEECH/12/149, 5 March, Bratislava, http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm , accessed 6 April 2015.

Lazer, D. et al. (2014), "The parable of Google flu: Traps in big data analysis", *Science*, Vol. 343, 14 March, http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

Loshin, D. (2002), "Knowledge integrity: Data ownership", 8 June, www.datawarehouse.com/article/?articleid=3052, accessed 6 April 2015.

Loukides, M. (2014), "The backlash against big data, continued", *O'Reilly Radar*, 11 April, http://radar.oreilly.com/2014/04/the-backlash-against-big-data-continued-2.html, accessed 6 April 2015.

McNamee, R. (2009), "Obama needs to think bigger about infrastructure", Huffington Post, The Blog, 2 July, www.huffingtonpost.com/roger-mcnamee/obama-needs-to-think-bigg_b_156126.html, accessed 6 May 2015.

Merriam-Webster (2014), "Infrastructure", *Merriam-Webster.com,* 24 October, www.merriam-webster.com/dictionary/infrastructure, accessed 6 April 2015.

MGI (McKinsey Global Institute) (2013), "Open data: Unlocking innovation and performance with liquid information", McKinsey & Company, October, www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information, accessed 3 March 2014.

Moody, D. and P. Walsh (1999), "Measuring the value of information: An asset valuation approach", Seventh European Conference on Information Systems (ECIS'99), Copenhagen Business School, http://si.deis.unical.it/zumpano/2004-2005/PSI/lezione2/ValueOfInformation.pdf, accessed 6 April 2015.

Newman, N. (2013), "Taking on Google's monopoly means regulating its control of user data", *Huffington Post*, The Blog, 24 September, www.huffingtonpost.com/nathan-newman/taking-on-googles-monopol_b_3980799.html, accessed 6 April 2015.

NRC (1987), *Infrastructure for the 21st Century: Framework for a Research Agenda*, Committee on Infrastructure Innovation, National Research Council, National Academy Press, Washington, DC..

O'Neil, C. (2013a), "K-nearest neighbors: Dangerously simple", 4 April, *Mathbabe*, http://mathbabe.org/2013/04/04/k-nearest-neighbors-dangerously-simple/, accessed 6 April 2015.

OECD (2014), *Addressing the Tax Challenges of the Digital Economy*, OECD/G20 Base Erosion and Profit Shifting Project, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264218789-en.

OECD (2013a), "Knowledge networks and markets", OECD Science, Technology and Industry Policy Papers, No. 7, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k44wzw9q5zv-en.

OECD (2013b), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, 11 July, C(2013)79, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2013c), "New data for understanding the human condition: International perspectives", OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences, February, OECD Publishing, Paris, www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf.

OECD (2012), "Exploring the economics of personal data: A survey of methodologies for measuring monetary value", *OECD Digital Economy Papers*, No. 220, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k486qtxldmq-en.

OECD (2011), Quality Framework and Guidelines for OECD Statistical Activities, OECD Publishing, Paris, 17 January, http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs%282011%291, accessed 6 April 2015.

OECD (2010), *The OECD Innovation Strategy: Getting a Head Start on Tomorrow*, OECD Publishing, Paris.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, 30 April 2008, [C(2008)36], OECD Publishing, Paris, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2006a), "Quasi Option Value", in OECD, *Cost-Benefit Analysis and the Environment: Recent Developments*, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264010055-11-en.

OECD (2006b), OECD Recommendation of the Council concerning Access to Research Data from Public Funding, 14 December 2006, [C(2006)184], OECD Publishing, Paris.

OECD (2005), *Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris, www.oecd.org/sti/sci-tech/38500813.pdf.

OECD (2004), Declaration on Access to Research Data from Public Funding, OECD Publishing, Paris, www.oecd.org/science/sci-tech/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm, accessed 6 April 2015.

Oppenheim, C., J. Stenson and R. Wilson (2004), "Studies on information as an asset III: Views of information professionals", *Journal of Information Science*, Vol. 30, No. 2, pp. 181-90.

Rochet, J.-C. and J. Tirole (2006), "Two-sided markets: A progress report", RAND Journal of Economics, RAND Corporation, Vol. 37, No. 3, pp. 645-67, http://ideas.repec.org/a/bla/randje/v37y2006i3p645-667.html.

Rose, C. (1986), "The comedy of the commons: Custom, commerce, and inherently public property", Faculty Scholarship Series, Paper 1828, http://digitalcommons.law.yale.edu/fss_papers/1828, accessed 6 April 2015.

Rotella, P. (2012), "Is data the new oil?", *Forbes*, 2 April, www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/, accessed 6 April 2015.

Savitz, E. (2011), "The new factors of production and the rise of data-driven applications", *Forbes*, www.forbes.com/sites/ciocentral/2011/10/31/the-new-factors-of-production-and-the-rise-of-data-driven-applications/, accessed 13 December 2013.

Schwartz, J. (2000), "Intel exec calls for e-commerce tax", *The Washington Post*, 6 June.

Scofield, M. (1998), "Issues of data ownership", www.information-management.com/issues/19981101/296-1.html, accessed 6 April 2015.

Shapiro, C. and H.R. Varian (1999), *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston, MA.

Steinmueller, W.E. (1996), "The US software industry: An analysis and interpretative history", in David C. Mowery (ed.), *The International Computer Software Industry*, Oxford University Press.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of government in a digital age", Computer and Communications Industry Association, October, www.ccianet.org/CCIA/files/ccLibraryFiles/Filename/000000000086/govtcomp_report.pdf, accessed 10 October 2013.

Thorp, J. (2012), "Big data is not the new oil", *HBR Blog Network*, 30 November, http://blogs.hbr.org/cs/2012/11/data_humans_and_the_new_oil.html, accessed 6 April 2015.

Trotter, F. (2012), "Who owns patient data? Look inside health data access and you'll see why 'ownership' is inadequate for patient information", O'Reilly Strata, 6 June, http://strata.oreilly.com/2012/06/patient-data-ownership-access.html, accessed 6 April 2015.

Ubaldi, B. (2013), "Open government data: Towards empirical analysis of open government data initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

UN (2008), *System of Nation Accounts 2008*, United Nations, http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf, accessed 6 April 2015.

UN Global Pulse (2012), "Big data for development: Opportunities & challenges", United Nations Global Pulse, May, www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf, accessed 6 April 2015.

United Kingdom Cabinet Office (2012), "Open data white paper: Unleashing the potential", June, http://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf, accessed 6 April 2015.

Zax, D. (2011), "Is personal data the new currency?", *MIT Technology Review*, 30 November, www.technologyreview.com/view/426235/is-personal-data-the-new-currency/, accessed 6 April 2015.

Chapter 5

Building trust for data-driven innovation

This chapter provides an overview of emerging trust issues raised by the increasing use of data-intensive applications that impact individuals in their commercial, social and citizen interactions. Security issues are addressed first, with an examination of the traditional approach and its inherent limitations. Comparisons are then made with current digital security risk management, which views risks as the possible detrimental consequences *for the objectives of, or benefits expected from, the data value cycle. The point is made that a certain level of risk has always to be accepted for the value cycle to provide some benefit – raising the question of who decides that level. The discussion then takes up privacy protection. Practical means for preventing information discovery are enumerated, and the dangers of information asymmetry, data-driven discrimination, and unanticipated uses of consumer data addressed. Attention then turns to potential policy approaches to help in addressing the issues raised.*

It takes years to build up trust, and only seconds to destroy it (unknown author).

"Ginny!" said Mr. Weasley, flabbergasted. "Haven't I taught you anything? What have I always told you? Never trust anything that can think for itself if you can't see where it keeps its brain?" (Rowling, 1998, *Harry Potter and the Chamber of Secrets*)

Critical to reaping the substantial economic benefits of data-driven innovation (DDI) – as well as to realising the full social and cultural potential of that innovation – is the key element of trust. Trust is a complex issue, and yet there is consensus that it plays a central if not vital role in social and economic interactions and institutions (Putnam et al., 1993; Morrone, et al., 2009; OECD, 2011a). In reducing transaction costs and frictions, trust generates efficiency gains, and is considered by some to be a determinant of economic growth, development, and well-being. The OECD (2011a) provides quantitative evidence that high country trust is strongly associated with high household income levels.

In relation to the digital economy, the main components of trust are security and privacy protection for individual citizens and consumers. DDI relies on an intricate, hyper-connected information and communication technology (ICT) environment in which security threats have changed in both scale and kind. Security measures aim to address these challenges to establish the trust needed for economic activities to take place. But they can also inhibit economic and social development, by reducing innovation and productivity. The digital security risk management approach described below is a way forward: it helps address security-related uncertainties in a manner that fosters DDI. Tackling the issues raised by the second dimension of trust – the protection of personal data – is less straightforward than addressing security, as will be seen below. In the context of DDI, the consumer protection issues relate primarily to the collection and use of consumer data and are treated in common with the more general analysis of privacy.

5.1. Security for data-driven innovation

Given that decision making is becoming increasingly data-driven and automated, and the expected benefits of such decision making and of data-driven knowledge creation are growing, it has become essential to address digital security The digital assets of businesses, individuals and governments face various types of threats from a growing number of sources. These include organised crime groups, "hacktivists", foreign governments, terrorists, individual "hackers" – and sometimes, business competitors. A range of techniques are deployed – some basic, others extremely sophisticated – to target valuable digital assets. There are in addition the non-intentional digital threats, such as hardware failure and natural disasters. Whether intentional or not, digital threats can disrupt the functioning of systems and networks and damage the economic and social activities that rely on the confidentiality, integrity and availability of data and information.

Analysis of a new generation of national cybersecurity strategies in OECD countries shows that cybersecurity policy making has reached a "turning point" and has "become a national policy priority" (OECD, 2012a). Yet many stakeholders continue to adopt a traditional security approach that not only falls short of appropriately protecting assets in the current digital environment, but also is likely to stifle innovation and growth. This traditional approach is introduced below, prior to a discussion of the digital security risk management approach promoted by the OECD.[1]

The traditional security approach

The traditional security approach, as conveyed by the dictionary definition of the term "security" (Online Etymology Dictionary, 2014), aims to create a secure environment that is "free from danger or risk of loss". Knowing that any threat will be eliminated or neutralised by security measures, users of the digital environment can then trust it and carry out their economic and social activities without being concerned.

In more precise terms, this approach aims to create a digital environment secure from threats that can undermine the availability, integrity and/or confidentiality of information or information systems – the AIC triad.[2] Availability is the accessibility and usability of data upon demand by an authorised entity. Integrity is the protection of data quality in terms of accuracy and completeness. Confidentiality refers to the prevention of data disclosure to unauthorised individuals, entities or processes.

To preserve each of these dimensions, security experts put in place security measures, sometimes also called security "controls", "mechanisms" or "safeguards". They are generally based on technologies (e.g. firewalls, anti-virus protection, encryption), people (e.g. training, assigning responsibilities) and processes (e.g. backup procedures, password policies). Thus it is possible for many security measures to be selected and implemented; however, since resources are limited, security experts often have to decide which measures to place where in a system for security to be the most effective. These choices are based on analysis of the likelihood of threats exploiting vulnerabilities and undermining one or more dimensions of the AIC triad.

Under this traditional model, once security measures are in place, the system and the valuable data it contains (i.e. the environment) are deemed protected. Security measures form a perimeter around the protected assets to secure them. As economic and social activities require limited protection within the walled perimeter, the effort is focused on the height of the walls, as well as the number of gates and guards controlling entry and exit. Finally, since security focuses on the protection of the digital environment, a key characteristic of the traditional security approach is that the primary responsibility for security generally rests with the party responsible for the provision of that environment: the IT department.

Limitations of the traditional security approach in a data-intensive environment

Data-driven innovation (or data-intensive economic and social activities) relies greatly on the digital environment. However, this environment must have certain characteristics to be conducive to DDI: it must be open and interconnected, as well as flexible. It must also host a massive volume of data of considerable diversity. Unfortunately, these interrelated characteristics increase the complexity of security management to a point where the traditional security approach cannot scale up.

Openness is essential and interconnectedness is blurring the perimeter

First, data-driven innovation leverages the fundamentally open and interconnected nature of information systems and networks, qualities enabled by the generalisation of Internet technologies in the second half of the 1990s. It depends on the capacity to exchange data flows easily, flexibly and cheaply with a potentially unlimited number of partners outside the perimeter.

The traditional closed security perimeter approach is thus an obstacle to the development of data-driven innovation. The idea that, for security reasons, a system should be kept closed by default and open only by exception belongs to the past, when information technologies were not designed for interoperability and when their contribution to economic and social progress depended less on the free flow of data. In the pre-Internet era, information systems were inherently isolated, designed and operated with a clear and closed perimeter by default. Interconnecting them required an expensive add-on to be developed on a case-by-case basis. However, since the mid-1990s, this "siloed" world has progressively disappeared. Information systems are now designed to

be open and interconnected by default, without additional cost, and this characteristic has become the main driver for using ICTs to drive productivity, innovation and growth. It has in fact become complicated and expensive to close information systems, both in terms of the security measures needed to reduce interconnectedness and – most importantly – because limiting interconnectedness also inhibits ICT from enabling any of those gains. Closing these systems will moreover provide only an illusion of security.

The very concept of a perimeter in any case becomes blurred in an environment where the number of gateways to the outside digital world increases exponentially, making systematic and comprehensive control of inputs and outputs equally illusory. It is challenging to define a perimeter whose length may cross the boundaries of the organisation and national jurisdictions, and perhaps even extend to the whole of the Internet. Trends such as cloud computing, mobility, "bring your own device", machine-to-machine communication (M2M) and the "Internet of Things" (i.e. the interconnection of physical objects over the Internet) are firm evidence of the dissolution of boundaries for information systems and networks. Future trends unknown now will likely continue to expand this landscape. Thus, although a system's perimeter remains a potential location to deploy local security measures, relying on its robustness while limiting its openness would be both ineffective from a security perspective, and economically counterproductive.

The traditional approach reduces the flexibility of the environment

Second, DDI relies on the capacity to exploit the dynamic nature of the digital environment – rapidly connecting, matching and analysing what was previously not related in order to create new assets. In contrast, the traditional security approach is meant to protect clearly defined tangible and intangible assets, including information systems, networks, data and information. Changes in the digital environment or its use are likely to require the reorganisation of security measures and are therefore not welcome from a traditional security perspective. Traditional security is static in nature and fails to address rapid change in the system and its use. If the economic benefits of DDI result from the dynamic nature of the digital environment and its usage, traditional security is likely to become an obstacle to change and therefore an inhibitor for realising the full benefits.[3]

The volume and diversity of data increases complexity

Third, the growing volume and diversity of digitised data, DDI's key enablers, raise another challenge to traditional security. Traditional security can deal with increased volumes and diversity if the data are located within a defined perimeter and their processing is not subject to continuously unpredictible uses and flows. However, the uncertainty already introduced by the open and dynamic nature of data-driven innovation grows, sometimes exponentially, with these increases. Malicious elements are ubiquitous in complex systems – just as the natural environment is rife with viruses, bacteria and parasites (Forrest, Hofmeyr and Edwards, 2013) – and the complexity of the security equation is now multiplied by the scale, volume and diversity of the data stored and processed. This can be seen as an aggravation either of the potential threat (i.e. more data are likely to attract more malicious players and generate more errors) or of the potential vulnerabilities – or both.

These characteristics underline the main weakness of the traditional security approach, which is both binary (there is no middle ground between secure and insecure) and monolithic (the entire digital environment has to be secured for the approach to be effective). By adding an ever growing number of fast-changing variables, the potentially unlimited extension and openness of the perimeter and the ever growing and

unpredictible (i.e. innovative) usage of the environment for legitimate economic and social objectives put stress on the security paradigm. The approach can only operate at the cost of reducing the complexity and increasing stability, which will inevitably slow innovative usage and, ultimately, undermine the economic and social benefits of interoperable ICTs.

One may argue that to address these challenges, traditional security could be implemented in a more flexible way by focusing on those parts of the digital environment that are of more value to the organisation, and placing less emphasis elsewhere. However, the value is not really in the digital environment itself, or in the data, or in the information, but rather in the whole data value cycle (see Chapter 1 of this volume); that is what can generate economic and social benefits. More precisely, even when focusing on the data rather than on the information systems and networks, the fact that "information is context-dependent" (Chapter 4) makes it difficult to tailor the traditional security approach to the value of the data, because they are impossible to assess before their use.

From traditional security to digital security risk management

As noted above, the challenges to digital security are not new; they result from the mid-1990s shift of ICTs towards openness and interconnectedness by default, a trend that fed two decades of flourishing Internet-related innovation. In the early 2000s, application to ICT-related economic and social activities of risk management concepts and frameworks experienced in other areas such as industrial, health and environmental risks offered an alternative to the traditional digital security approach. The risk-based approach – which has been referred to with different terms and sometimes misleading but widespread expressions such as information security, information assurance and cybersecurity – requires a different culture, mindset and framework from traditional security. It redefines what should be protected, for what purpose, how it should be protected, and who should be responsible, with far-reaching consequences in terms of corporate governance applied to ICTs, and business management more generally. Surprisingly, however, while its application to ICTs is relatively recent, risk management is a common management and decision-making tool in business and industry.

The methodology of application is actually the same used by decision makers to address other risks. Nevertheless, it represents a significant paradigm shift in the way economic and social (or "business") decision makers,[4] security experts and ICT professionals often approach digital security threats. This section focuses on the two main changes required by digital security risk management: a different culture, and a different framework. The application of digital security risk management to digital activities was initially promoted in the 2002 *OECD Recommendation concerning Guidelines for the Security of Information Systems and Networks,* and is at the core of an ongoing process to revise this Recommendation.

From a culture of security to a culture of risk management

As explained in Chapter 1 of this volume, the value of data-intensive activities is not limited to the digital storage and processing of a large quantity of data ("big data"), but rather to the capacity to manage a data value cycle (see Figure 1.7 in Chapter 1). This cycle can transform the data into information and knowledge to feed more effective decision making and generate economic and social benefits through DDI. The objective of digital security risk management is therefore to *increase the likelihood of economic*

and social benefits from the data value cycle by minimising potential adverse effects of uncertainty related to the availability, integrity and confidentiality of the cycle (AIC triad). Unlike the traditional security approach, digital security risk management does not aim to create a secure digital environment to eliminate risk. Instead, it creates a framework to select proportionate and efficient AIC security measures in light of the benefits expected from the cycle. Therefore, it should be an integral part of the establishment and business use of the data value cycle, rather than merely a technical framework or a process separated from the business cycle.

The application of risk management to the use of ICT throughout the data value cycle requires decision makers and other key actors to understand the following fundamentals:

- Digital security risks are the possible detrimental *consequences* for the objectives of or benefits expected from the data value cycle that could result from uncertain events.[5] Such events are generally incidents resulting from the nexus of threats and vulnerabilities. In simple terms, risks are not the causes of problems but their economic and social consequences. Although it is important to understand the causes, digital security risk management focuses primarily on their potential economic and social consequences.
- To generate benefits, the data value cycle relies on *open, interconnected, dynamic and flexible ICTs*. In most contexts, these characteristics are essential to realising the benefits of data-driven innovation (must-have features). They are not optional add-ons (nice-to-have features) that can be simply dismissed or limited. Limiting them will directly impact the expected economic and social benefits of data-intensive activities.
- Because of this indispensable openness and interconnectedness, a degree of uncertainty is inevitable and must be accepted. The digital environment cannot be secured or made completely safe. Despite all the security measures that may be put into place, risk related to the use of ICTs cannot be completely eliminated. Thus *a certain level of risk has always to be accepted (i.e. taken)* for the value cycle to provide some benefit. This is often called "residual risk". The ability to manage risk is a critical success factor in DDI.

The risk management framework helps determine which security measures can reduce risk to an acceptable level[6] in light of the potential benefits, while recognising that the same measures impose constraints on the economic and social activities at stake. These interferences should be understood and balanced with the benefits before the measures are agreed upon and implemented. They can affect performance, cost, complexity and usability, in turn impacting on profitability and time to market. They can also impact on privacy. Thus digital security risk management is a disciplined systematic approach to achieve the right balance between insufficient security measures – i.e. where the benefits are undermined by an unacceptable level of risk – and too many security measures – i.e. where the benefits are inhibited by too much security.

That raises the key question of responsibility. Traditional security focuses on securing the digital environment. Therefore, in most cases, the party responsible for the provision of the environment takes responsibility for its security, and users of the environment do not have to be concerned with it. In contrast, from a digital security risk management perspective, responsibility cannot be delegated to a separate party. Instead, the allocation of responsibility follows three principles, all of which stem from management being an integral part of economic and social (i.e. "business") decision making.

First, as noted above, managing risk means accepting a certain level of risk – or, deciding not to accept it, and therefore not to realise the benefits. The primary responsibility for managing risk should mirror the responsibility for achieving the objectives and realising the benefits. Digital security risk is not an exception to this general management principle.

Second, in complex data-driven innovation activities, it is likely that only one or a very limited number of actors will be responsible for realisation of the overarching expected benefits from the data value cycle; many other actors will each have a role to contribute, at their level, to achievement of these objectives. Among them, some will ensure that the content of the data value cycle generates the expected benefits ("business" or economic actors); others will provide the optimal digital environment to support the operation of the cycle (IT actors).[7] The distribution of responsibility for digital security risk management should therefore reflect this distribution of roles, and appropriate delegations of responsibility and authority to act should be established. Since benefits and risks are two sides of the same coin, the ultimate owner of the benefits should also ultimately own the risk. One consequence is that the primary responsibility should not be fully delegated to an IT actor who could jeopardise the economic performance of the data value cycle.

Third, DDI relies on a chain of interdependent links; to some degree, each part of the value cycle is dependent on and impacts the others. In the cycle's operation, the responsible actors cannot be isolated from one another: they form a holistic data ecosystem. The complexity of interdependencies within this ecosystem can be very high, considering for example that some elements or sub-elements of the cycle can operate across different organisations, and even across jurisdictions. Therefore, a clear mapping of roles and responsibilities is necessary. Further, digital threats can propagate extremely rapidly from one link to the other, quickly contaminating the entire cycle. The consequences of threats exploiting vulnerabilities can escalate both within the cycle and beyond, to affect other economic and social activities that depend on the cycle. For that reason, it is indispensable to establish a robust collaboration and co-operation culture supporting good communications among the business actors and the IT actors, and between the two groups. Thus to manage risk in the part of the cycle where Party A has a responsibility, a security measure may be more effective in a part under the responsibility of Party B. Similarly, security measures protecting the part of the cycle under one's responsibility may create risk in or undermine the effectiveness of another part. Considering this degree of complexity, it could be good practice to assign responsibility to someone for ensuring co-operation and collaboration among all the cycle's actors.

Possible good practices include the following examples (see also Box 5.1):

- There could be a clear rule whereby if a risk management decision made at a particular stage in the cycle would affect another stage, then it should be made collectively among the group of decision makers with responsibility in the areas of possible impact.
- Digital security risk management decisions affecting the performance of the cycle as a whole should be made at the highest level of responsibility by the actor who is responsible for the benefits from the cycle and, as such, can understand the effect of such decisions on the cycle's objectives; assess the tolerance to risk (or "risk appetite") in light of the expected benefits; and set the appropriate acceptable level of risk.

Box 5.1. **An illustration of digital security risks**

Consider a large national or international discount supermarket chain selling mass products at a low margin and aiming to optimise its profits through high-volume sales. Data and analytics may support that objective by enabling the company to optimise its supply chain in respect to the expected demand of its customers. The company would establish a data value cycle to optimise the decision-making processes related to the chain's purchase, logistics, and marketing activities including (e.g.) price and discounts, product placement and advertising. However, the supermarket chain would not control all the data sources, as it would rely on many third parties outside its control (e.g. data brokers) to feed the cycle or even to operate some parts of it (e.g. cloud computing providers).

In this context, the AIC triad is key to data-driven innovation for the company. A breach of availability at one stage could stop or slow the ability of the company to maximise its profit margins, leading to economic losses or lost opportunities. Integrity is also a key factor. Accurate economic decisions from the cycle rely on accurate data and information flowing through the cycle and accurate analytics. Wrong decision making directly affecting the company's profitability and competitiveness could result from illegitimately modified data and information at any stage of the cycle, or from corrupted analytics. Finally, confidentiality relates to protecting the company's market position. For example, a competitor could greatly benefit from disclosure of the underlying analytic algorithms or from obtaining access to the company knowledge base, which is essential to key economic decisions and part of the company's competitive advantage. A breach of confidentiality at the value cycle's "big data" stage, i.e. data storage, could result in legal privacy challenges damaging the company's reputation and consumer trust, in addition to potential financial loss from lawsuits.

In assigning decisions, one should also clearly distinguish between those relating to provision of the digital environment (i.e. IT), which do not have a negative impact on the performance of the data value cycle (such as updates and upgrades), from those that can impact its performance and deliverables. The former should be left to the IT professionals; the latter should be first reviewed by the actors responsible for realising the objective of the value cycle, since they are best placed to understand the potential consequences for these objectives. All responsible actors should, however, take advantage of the essential role played by IT professionals in managing technical security measures and informing other decision makers about threats and vulnerabilities. Good co-operation and dialogue are essential.

Since the cycle may bring together professionals with different skills, cultures and perspectives, ensuring a high degree of collaboration will likely prove a management challenge. Establishment from the outset of a common culture of risk management would help transcend differences across various groups of actors and increase team spirit and project coherence. The creation of such a culture is a critical success factor. The multiplication of stakeholders involved in operating the data value cycle from outside the boundaries of the organisation, such as cloud providers and other outsourced services, further increases the challenge.

As noted above, many different actors with different roles will operate the data value chain and reflect different background and styles of management which need to be respected. With regard to these many different actors, digital security risk management should respect fundamental rights and values, such as privacy, as well as the legitimate interests of others. Privacy is a case where the ethics and interests of those who establish, use and benefit from the data value cycle may not be fully aligned with those whose personal data are processed. Mechanisms to reconcile conflicting interests are needed.

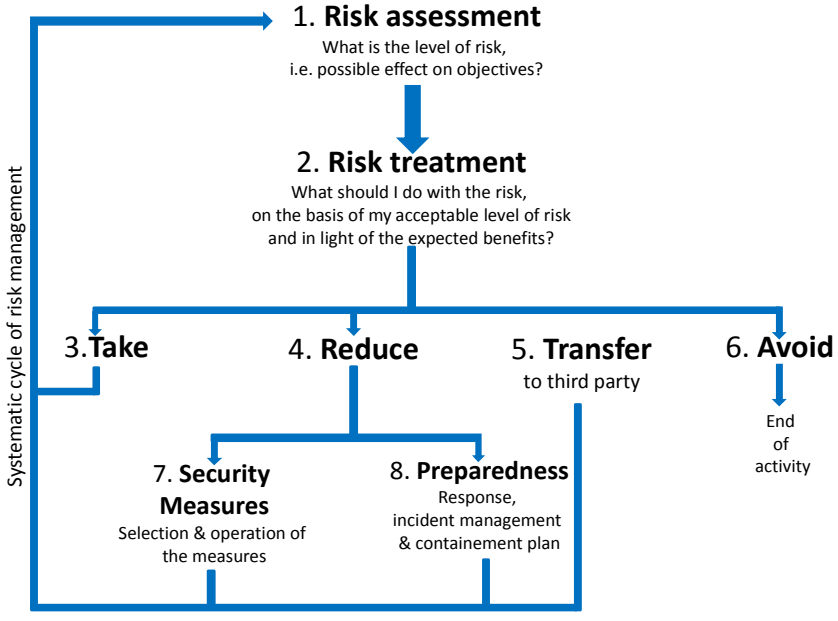
The owner of the cycle should be responsible to ensure that across each of its stages, digital security risk management is accomplished in accordance with the rights and values, regulation and culture of all parties, including parties considered "external" to the cycle but affected by it.

In summary, a culture of digital security risk management is essential to increase the likelihood of success of data-driven innovation projects. It relies on a solid understanding of risk management; the alignment of responsibility for managing risk with the responsibility to realise the objective; a robust collaboration and co-operation framework; and respect for the fundamental values and legitimate interests of others.

From static perimeter security to risk management cycle

While the data value cycle in Figure 1.7 (Chapter 1 of this volume) is a valid general representation of most data-intensive innovation activities, its complexity can vary considerably depending on the size of the organisation, the number of people and entities involved (stakeholders), and many other factors. In some cases, as noted above, such a value cycle can cross multiple organisations and regulatory jurisdictions, and consequently involve a multiplicity of different professional cultures and perspectives. The complexity of the efforts to generate benefits will be mirrored in the complexity of managing digital security risk. The response to a higher degree of complexity is to establish a systematic framework for digital security risk management processes and weld it together with the data value cycle. Doing so is essential to enable risk management to scale up, as well as for auditing and accountability reasons. Figure 5.1 represents the digital security risk management cycle.

Figure 5.1. **Digital security risk management cycle**



As with the traditional security approach, security here is related to the likelihood of threats exploiting vulnerabilities to undermine the AIC triad, and the security measures available are the same in both approaches. However, their selection and application results from a completely different process that starts with the assessment of risk (Step 1 in Figure 5.1) and its treatment (Step 2), i.e. the determination of whether to take it as it is

(Step 3), reduce it (Step 4), transfer it to someone else (e.g. through contract, insurance or other legal agreement) (Step 5) or avoid it by not carrying out the activity (Step 6). If one decides to reduce the risk, the risk assessment helps determine which security measures should be selected and applied where and when, in light of the consequences of uncertain events on the economic and social objectives (Step 7). The primary criterion determining selection and application is the acceptable level of risk to the economic and social activities at stake – not just the likelihood of a threat that can exploit a vulnerability to create harm. This process provides shades of grey that enable the systematic protection of assets in proportion to their value, and therefore enable the protection to scale to the size and complexity of the value cycle. Finally, residual risk cannot be ignored. A preparedness plan (Step 8) should also be established to limit and manage the consequences of incidents when they occur and reduce the potential of escalation. Finally, since the dynamic nature of the cycle is key to DDI, the risk assessment and treatment, selection of security measures and incident management plan should likewise be operated as an ongoing cycle.

In the complex context of DDI, risk should be assessed at several levels: the level of the data value cycle as a whole; the level of each stage of the data value cycle (as represented in Figure 1.7); the level of each sub-process within each of these stages, and so on until the degree of uncertainty is considered sufficiently understood by the owner of the benefits and risk. In reality, it is the aggregation of risk assessments at the lower level that will feed the risk assessment at higher levels up to the highest one. This will produce an overarching decision-making tool in the form of a risk matrix surrounding the data value cycle and encompassing the whole data ecosystem.

5.2. Privacy protection for data-driven innovation

Unlike the analysis of digital security risks, which are tied to the economic and social objectives of the data-driven activity in question, analysis of privacy issues is oriented around the impact on individuals and society, whose interests may not always fall directly within those objectives. There are clear areas of overlap between security and privacy, including the security safeguards principle of the OECD (2013a) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) and the utility of the risk-based approach. But there are also important differences in considering privacy issues, some of which touch on fundamental values for individuals and society. After identifying certain privacy-related impacts, the discussion turns to a number of policy approaches that could offer more effective privacy protection.

Possible privacy-related impacts of data-driven innovation

Trends in data collection, analysis and use described in other chapters of this volume are transforming organisational practice across a number of business and government sectors, and could extend to many more. At the same time, these trends are raising questions about whether policies, laws and norms are able to protect privacy and other social values, such as individual liberties, that are essential for user trust in the digital economy. The issues described below may not be truly novel, but taken in combination they pose important and fresh questions about how to ensure that DDI will be deployed to the benefit of individuals, whether as consumers, social actors or citizens.

Each step of the data value cycle (Figure 1.7) on which DDI relies can raise privacy concerns. Step 1 is the initial data collection, which is becoming increasingly

comprehensive, diminishing an individual's private space. Step 2 is the massive storage of data, which increases the potential of data theft or misuse by malicious actors and other consequences of a data security breach, the risks of which may not be easy to ascertain. Steps 3 and 4 involve inferences of information and knowledge[8] enabled by data analytics, a tool that extracts information from data by revealing the context in which the data are embedded and their organisation and structure.[9] Analytics often goes well beyond the data knowingly provided by a data subject, diminishing an individual's control and creating information asymmetry. Finally, data-driven decision making (Step 5) can lead to a real-world discriminatory impact on individuals and other harms.

Comprehensive data collection and the loss of private space

Many commercial and social activities, whether conducted in public or in private, leave behind some form of digital trace. A growing number of entities, such as online retailers, Internet service providers (ISPs), operating systems, browsers, social media and search engines, financial service providers (i.e. banks, credit card companies, etc.) and mobile operators have the capability to collect vast amounts of this data. Such data collection may be limited to a specific context or transaction, but usually spans a wide range of economic social activities.

Some of the data collected are knowingly and willingly provided by the consumer, and are often essential to the completion of an online commercial transaction. Behavioural advertising, by contrast, relies on the online tracking of consumers and the collection and analysis of related personal information in order to provide them with advertising tailored to their expected needs and interests. Another example is geolocation data from mobile devices, which on the one hand can be used to improve the location-based services on which many rely today, but at the same time leaves a trail of an individual's daily routines and movements, which are increasingly used for other services including for process improvements. According to a recent survey, two-thirds of device owners in the United States have no idea who has access to data from their devices or how it is used (Intel, 2014).

Other types of data are not collected directly from the consumer. Data brokers, for example, collect and aggregate personal data regarding individuals with whom they have no direct interaction, in order to offer a variety of services to third parties, such as employment background checks, localisation services and identity verification (FTC, 2014). Government and private sector researchers are increasingly using health data to evaluate outcomes, identify drug interactions, and push the boundaries of predictive medicine (OECD, 2013d). In education, as technological tools become commonplace in schools, data collection, retention, and analysis are becoming increasingly systematic, revealing greater insights into student and teacher performances (*New York Times*, 2014). The possibility of improving educational performance is evident, but so too is the challenge of introducing policies and processes for protecting sensitive student data.

While the use of data collected about individuals can benefit organisations, individuals and society, the breadth and scale of current data collection practices has given rise to concerns on the part of data subjects, and these have both social and economic ramifications. First, if citizens believe that they are being watched or monitored with respect to their online activities in ways they consider inappropriate or unfair, they may feel less free to participate in the discussion of controversial subjects; this can lead to a type of self-censorship that may undermine civil discourse and engagement. Indeed, one study indicates that citizens in several OECD countries have begun to self-censor with

respect to search terms entered into search engines, based on a widespread public belief that their use of search terms is being inappropriately monitored (Marthews and Tucker, 2014). Second, individuals may feel that certain forms of commercial data collection are inappropriate or unfair, particularly when they are being carried out without their knowledge or consent. Concerned that storage of a massive aggregation of personal data is more vulnerable to privacy violation, including through inappropriate reuse, individuals may simply forego certain online activities.

Inference and the loss of control

There are a number of means that individuals use to protect their own privacy (see Box 5.2). Intuitively, the most obvious way is to withhold or conceal information relating to them. However, the ubiquitous nature of ICTs, coupled with technological advances in data analytics, makes it increasingly easy to generate inferences about individuals from data collected in commercial or social contexts, even if these individuals never directly shared this information with anyone.

Box 5.2. **Practical means for preventing information discovery**

Data analytics extracts information from data by revealing the context in which the data are embedded, its organisation and structure (see Chapter 3 of this volume). There exist a number of practical means for preventing or significantly increasing the cost of extracting the information embedded in the data through data analytics, though these may adversely affect data utility. Examples follow.

Reduced collection – A reduction in data collection can be considered the strongest means for preventing information extraction, because where no data are collected, no information can be extracted. Data subjects can withhold or decline to provide data. Data controllers can practice data minimisation. As Pfitzmann and Hansen (2010) have highlighted, data minimisation "is the only generic strategy [misinformation or disinformation aside] to enable unlinkability, since all correct personal data provide some linkability".

Cryptography – Cryptography is a practice that "embodies principles, means, and methods for the transformation of data in order to hide its information content, establish its authenticity, prevent its undetected modification, prevent its repudiation, and/or prevent its unauthorised use" (OECD, 1977). It is a key technological means to provide security for data in information and communications systems. Cryptography can be used to protect the confidentiality of data, such as financial or personal data, whether that data is in storage or in transit. Cryptography can also be used to verify the integrity of data by revealing whether data have been altered and identifying the person or device that sent it.

De-identification – This term covers a range of practices ranging from anonymisation to pseudonymisation. These practices share a common aim of preventing the extraction of identifying attributes (i.e. re-identification), or at least significantly increasing the costs of re-identification. *Anonymisation* is a process in which an entity's identifying information is excluded or masked so that the entity's identity cannot be, or becomes too costly to be, reconstructed (Pfitzmann and Hansen, 2010; Mivule, 2013). Some research suggests that when linked with other data, most anonymised data can be de-anonymised – that is, the identifying information can be reconstructed (Narayanan and Shmatikov, 2006; Ohm, 2009).[1] For many applications, however, some kind of identifier is needed, and having complete anonymity would prevent any useful two-way communication and transaction. *Pseudonymisation* is therefore used, whereby the most identifying attributes (i.e. identifiers) within a data record are replaced by unique artificial identifiers (i.e. pseudonyms).

Box 5.2. **Practical means for preventing information discovery (cont.)**

Unlinkability and functional separation – Unlinkability results from processes to ensure that data processors cannot distinguish whether items of interest are related or not (Pfitzmann and Hansen, 2010). According to ISO (Pfitzmann and Hansen, 2010), unlinkability "ensures that a user may make multiple uses of resources or services without others being able to link these uses together". De-identification is a means to enable unlinkability, but cannot guarantee it. Other technical means include functional separation and distribution (decentralisation).

Noise addition and disinformation – The addition of "noise" to a data set allows analysis based on the complete data set to remain significant while masking sensitive data attributes. Finding the right balance that protects privacy while minimising the costs to data utility is a challenge (Mivule, 2013). Disinformation is false or inaccurate information spread intentionally to mislead. Noise addition techniques are considered promising means to help protect privacy and confidentiality in databases, while keeping all data sets statistically close to the original data sets. Work on "differential privacy" is one example (Dwork and Roth, 2014).

1. Narayanan and Shmatikov (2007), for example, have used the "anonymous" data set released as part of the first Netflix prize to demonstrate how the authors could correlate Netflix's list of movie rentals with reviews posted on the Internet Movie Database (IMDb). This let them identify individual renters, and gave the authors access to their complete rental histories (Warden, 2011). The view that de-identification does not work has been challenged, however – see for example Cavoukian and Castro, 2014.

As highlighted in Chapter 3 of this volume, data analytics extracts information from data by revealing the context in which the data are embedded, including patterns, correlations among facts, interactions among entities, and relations among concepts (Merelli and Rasetti, 2013). Thus, data analytics enables the "discovery" of information even if there was no prior record of such information. Data analytics is not a new phenomenon. However, as the volume and variety of available data sets increase, as well as the capacity to link different data sets, so does the ability to derive further information from these data. Advances in analytics now make it possible to infer sensitive information from data that may appear trivial at first, such as past purchase behaviour or electricity consumption. The IoT will likely accelerate this trend, generating a large number of diverse but interlinkable data sets that relate to economic and social activities.

Traditionally, many privacy violations have involved disclosure of personal information beyond the envisaged recipients or beyond the purposes set out for the collection and processing of the data. However, active disclosure or new secondary uses are not a prerequisite for the inference of personal characteristics. Once linked with sufficient other information, data analysts can predict, with varying degrees of certainty, the likelihood that an individual will possess certain characteristics, building a profile. This knowledge can be used for legitimate purposes, but it can also be used in ways that individuals do not desire or expect, or which adversely affect them – for example, when it results in unfair discrimination. There is a risk that the inferences may not be accurate, but even where correct, the inferences produce possibilities for tactless, harmful or discriminatory use of data outside the individual's control (see Chapter 3 on the limits of data analytics).

Creation of the knowledge base and information asymmetry

In spite of the wealth of information now available to consumers via the Internet, businesses in many respects retain more information than consumers about the features and quality of the products they sell, the contract terms and conditions, the associated production and distribution costs, the availability of competing alternatives, and so forth. Data analytics can indeed lead to valuable insights for those parties employing them. In

the credit-reporting context, for example, the very purpose of data analytics is to reduce information asymmetries by making a debtor's credit history available to potential creditors (World Bank, 2011). Certain forms of data analytics, however, entail a risk of exacerbating the information imbalance – in certain cases dramatically (Schermer, 2011) – and in ways potentially incompatible with broader societal values.

For example, the use of data analytics may increase the ability of governments and businesses to influence and persuade individual citizens and consumers. As described earlier, companies are often capable of targeting a particular consumer with advertisements across a range of websites and apps. This capacity has also been used for political campaigns (Hurwitz, 2012; Rutenberg, 2013; Nickerson and Rogers, 2014). Tailoring advertisements to the interests of consumers may benefit both the company and the consumer. However, concerns have been raised that the information inferred through data analytics may also facilitate aggressive or predatory marketing practices, whereby a company exploits the vulnerabilities of consumers in a way that induces them to purchase goods or services that they would not otherwise have bought. In the context of government-citizen relationships in areas such as health and education, the same data-driven approaches that bring improved government delivery of services to citizens can also expand government power (EOP, 2014, p. 22).

Behavioural factors (e.g. the tendency of individuals to focus on short-term benefits and costs, their tendency to automatically accept defaults set by organisations, and their overconfidence) can lead individuals to make poor and sometimes costly decisions (OECD, 2010). The issue of information asymmetry can be further exacerbated by the limited transparency of data analytics, particularly with respect to data controllers who have no direct interaction with consumers. Many individuals may either be unaware that data analytics are affecting the marketing and delivery of goods and services they are being offered, or have considerable difficulty ascertaining how exactly analytics are being used to influence them or to determine the offers they see online. Individuals are becoming more transparent to organisations, but it is not clear that there is a parallel advance in the transparency of the data-processing practices of organisations to individuals.

Data-driven decision making, discrimination and other societal values

Data analytics can be used to provide new insights into human behaviour and societal and macroeconomic trends. In their search for competitive advantage, private actors increasingly rely on the predictive capabilities of data analytics. While these predictive analyses may result in greater efficiencies, they may also perpetuate existing stereotypes, limiting an individual's ability to escape the impact of pre-existing socio-economic indicators.

Classification based on attributes is at the heart of many forms of data analytics associated with profiling, an activity defined in the section "The pervasive power of data analytics" in Chapter 3 of this volume. Through data mining techniques that extract information patterns from data sets, a data analyst can discover patterns and relationships among different data objects, which in turn allow for increased differentiation. A well-known example in this regard is consumer segmentation: individual consumers are categorised among different behavioural or socio-economic profiles on the basis of observed or inferred attributes.

One form of differentiation among consumers that has recently been gaining attention is known as "price discrimination" (also referred to as "differential", "personalised" or "dynamic" pricing). Price discrimination is traditionally defined as firms' sale of the same good to different customers at different prices, even though the cost of producing for the

two customers is the same (OECD, 2002). This can occur directly, where each consumer is charged based on his or her willingness to pay, or indirectly, through volume discounts or discounts for groups such as students or the elderly (OFT, 2013).

Price discrimination is facilitated by data analytics in a number of ways. For example, by analysing a consumer's behaviour over a certain period, vendors can obtain a strong indication of future purchasing habits, which allows them to set their prices accordingly. Similarly, data analytics can allow vendors to differentiate among customers with different degrees of willingness to pay, by steering them towards different sets of products when they search within a product category (Valentino-Devries, 2012).

Proponents defend price discrimination practices on the grounds of efficiency and an ability to increase aggregate economic welfare. Opponents argue that such practices are unfair, violating notions of equality among consumers and exacerbating existing information asymmetries (see Box 5.3). Detailed consumer profiles enable vendors to obtain a strong indication of a consumer's demand curve and reserve price. The consumer, however, typically has no idea of a vendor's reserve value, and is disadvantaged as a result. While a consumer may retain the ability to shop elsewhere, the transactional costs can add up. There are also transparency issues, as consumers may be unaware that the prices they see are determined in part by their personal data collected in the past.

Box 5.3. **Consumer reaction to price discrimination**

- According to a survey carried out from January 2004 to May 2005, 87% of Americans surveyed strongly object to or would be bothered by the practice of online stores charging people different prices for the same products based on information collected about their shopping habits (Turow, Feldman and Meltzer, 2005).
- A number of studies suggest that consumers perceive personalised pricing as unfair (Garbarino and Lee, 2003; Levine, 2002; Hillman and Rachlinski, 2001; Odlyzko, 2003).
- Others suggest that consumers can accept price discrimination as long as all consumers have equal access to better prices and benefit from product choices (Cox, 2001; Dickson and Kalapurakal, 1994).
- Consumers tend to believe that coupons, rewards and discounts are fair practices (Narayanan, 2013).

Price discrimination is not new and, as a legal matter, is not generally considered to be an unfair commercial practice.[10] There is a long history of prices negotiated directly between businesses and consumers, or consumers being placed in different categories and charged accordingly. What is new is the potential that analytics creates to systematise personalised pricing. Anecdotal evidence suggests that the practice is not yet widespread (EOP, 2015). But if it were to become a more commonplace activity, how would consumers react?

While differentiating on the basis of price may be the most overt type of discrimination, analytics can also enable personalised treatment in other dimensions of the customer-business relationship. Customer service calls, complaint handling and many other interactions can be tailored to the specific customer. Consideration may be called for as to whether there are limits beyond which differential treatment of consumers should be considered a form of discrimination and discouraged.

Certain uses of data analytics may have additional and more serious implications for individuals, for example by affecting their ability to secure employment, insurance or credit. Indeed, this is the precise purpose of data analytics in credit reporting systems, to

support "unbiased credit decision-making … based on objective and correct data" and to "discipline debtor behaviour" by rewarding good credit history (World Bank, 2011, p. 23). Historically, credit has been granted on the basis of a credit officer's personal knowledge of the debtor. Ideally, advanced data analytics in credit reporting systems can empower consumers, enabling credit to those denied in the past due to some form of prejudice (e.g. assuming automatically that a low-income individual is always a bad debtor). At the same time, they also "raise the potential of encoding discrimination in automated decisions," in ways not fully transparent (EOP, 2014).

Segmentation and differentiation among individuals can yield important benefits to both organisations and individuals, ranging from well-targeted offers to credit to underserved communities, personalised medicine and improved fraud detection. Some forms of discrimination, however, are generally considered unethical or even illegal, such as differentiation on the basis of race, gender, ethnicity or disability (EOP, 2014, p. 64). Even when discrimination is based on less contentious characteristics (such as income level), there is still a risk of discrimination against certain social groups that might otherwise be protected. This may be the case, for example, where the characteristics used to differentiate are shared by a majority of individuals who belong to a particular social or racial group (Sweeney, 2013). Similarly, characteristics such as geographic location or postal code may serve as effective proxies to disguise what would otherwise be unlawful discrimination. In other words, even when the categories of differentiation that result from data analytics do not derive from prejudicial sources, they may nonetheless have a discriminatory effect against certain social groups in practice.

Discrimination may take other forms that run counter to societal values. To take an example related to fairness, consider a scenario in which an increasing number of people choose to collect and track data about their health, lifestyle, diet or even driving habits, and disclose them to their insurance company in exchange for discounted rates. Such an exchange may well serve their individual interests. But such practices may have social costs for others who choose not to share their data, for whatever reason. These individuals could eventually be charged higher rates than those who do share, or even be denied coverage – not because they represent a higher risk, but rather because they do not agree to participate in the profiling. Such scenarios may have policy implications with regard to fairness.

The links to freedom of speech and association are more difficult to discern, but of significant potential consequence. An environment where digital activity is systematically tracked and aggregated may create a "chilling effect" in which an individual curtails communications and activities in fear of uncertain but possibly adverse consequences (IWGDPT, 2014, p. 9). The widespread exchange of views with those whose opinions may differ may be undermined by what has been called the "filter bubble" effect when efforts to personalise news and other content narrow the range of views exposed to an individual (Pariser, 2012). More generally, data analytics can affect human decision making by shaping the behaviour of individuals, but it may also (unintentionally) alter individuals' preferences and, where the use of analytics undermines the values of those being influenced, set society on an irreversible transformative path (see Lessig, 1999; Frischmann, 2014).

Policy approaches for more effective privacy protection

Several responses can be identified for improving the effectiveness of privacy protections in the context of data-driven innovation. One set of initiatives is grouped under a heading of improving transparency, access and empowerment for individuals. A second area of focus is the promotion of responsible usage of personal data by

organisations. The promise of technologies used in the service of privacy protection has been long noted. Finally, the application of risk management to privacy protection is highlighted as providing another possible avenue.

These initiatives are not, of course, mutually exclusive and may best be deployed in combination with each other. Likewise, their implementation fits within the broader policy framework of an instrument such as the OECD Privacy Guidelines and the applicable legislation. A number of challenges have been noted in attempts to apply elements of the broader framework in terms of the scale of data use today. For example, the Report from the Expert Group that helped prepare revisions to the OECD Privacy Guidelines identified the role of consent, the role of the individual, the roles of purpose specification and use limitation, and the definition of personal data as raising issues for further study (OECD, 2013b).

Those challenges may not be surprising, given the historical context that shaped the elaboration of the principles. The predominant data processing model in the 1970s involved the direct provision of personal data from a data subject to a data controller. Although some examples of observed or inferred data can be found from that period, the basic model assumed an active role for the individual as a participant in the data collection process.

Today that assumption is challenged. The growth of the Internet, including the Internet of Things, has led to an explosion in observable data, with sensor-equipped smart devices poised to expand that category further. And the capacity to run analytics over unstructured data sets – which may be related to identified or identifiable individuals – is significantly expanding the category of inferred data and probability-based determinations. The context of processing addressed in this volume is focused to a much greater extent on data that are observed or inferred though sensors and analytics, which are growing at a much faster rate than user-contributed data (Abrams, 2014)

It should not be surprising, then, that the principles formulated for an active, engaged data subject are more challenging to apply where the personal data in question are generated at a distance from the subject. Nevertheless, attention to the areas identified below can help improve the effectiveness of privacy protection in a dynamic environment, where there is such flux in the scale, scope and value of personal data uses.

Transparency, access and empowerment

Promoting transparency and the rights to access and correction have been part of the OECD Privacy Guidelines since their initial adoption in 1980, and are incorporated into national laws around the world. Transparency and access have long been recognised as powerful tools against discrimination, as they help enable data subjects to ascertain the basis on which decisions are taken. The Council of Europe recommends that in some circumstances the transparency extend to include the logic underpinning the processing in the context of profiling (Council of Europe, 2010). However, many individuals today find it difficult to exercise these rights. There are a number of new initiatives aimed at rebalancing information asymmetries between individuals and organisations, and better enabling individuals to reap the benefit and value of their data. Governments are partnering with businesses to provide consumers with access to their personal data (including their own consumption and transaction data) in portable, electronic formats.

One element in a number of these initiatives is data portability, which allows users to more easily change data controllers by reducing switching costs, and enables them to analyse their own data for their own benefit by receiving it in a usable format. Data

portability not only promises to give individuals a key role in promoting the free flow of their personal data across organisations, thereby strengthening their participation in data-driven innovation processes; it is also seen as a means of increasing competition among providers of data-driven products.

In the United States for example, in 2011 the US National Science and Technology Council launched Smart Disclosure, an initiative aimed at providing consumers with access to data about products and companies as well as to their own data, in a secure, user-friendly and portable electronic format (NSTC, 2013). Other projects include the *Green Button* initiative, aimed at providing electricity customers with easy access to their energy usage data in a consumer-friendly and computer-friendly format. Recent efforts in the United States have focused on enhancing the transparency around the practices of data brokers (FTC, 2014) who have begun to respond with initiatives of their own.[11] In 2011, a government-backed initiative called *Midata* was launched in the United Kingdom to help individuals access their transaction and consumption data in the energy, finance, telecommunications and retail sectors. Under the programme, businesses are encouraged to provide their customers with their consumption and transaction data in a portable, preferably machine readable format. A similar initiative has been launched in France by Fing (Fondation Internet Nouvelle Génération), which provides a web-based platform MesInfos,[12] for consumers to access their financial, communication, health, insurance and energy data that are being held by businesses. Last, but not least, the right to data portability proposed by the EC in the current proposal for reform of their data protection legislation aims at stimulating innovation through more efficient and diversified use of personal data by allowing users "to give their data to third parties offering different value-added services" (EDPS, 2014).

These initiatives (discussed further in Chapter 4 of this volume), promise greater control to individuals wishing to make informed decisions and increase their trust in the data-intensive services that organisations seek to deliver. But such programmes may also bring significant costs, in terms of both developing and maintaining the mechanisms for enhanced data access and compliance with relevant regulations (Field Fisher Waterhouse, 2012). That raises the question about who should bear the costs for developing and maintaining these mechanisms.

Other initiatives aiming to address transparency issues for individuals include efforts to develop "multi-layered privacy notices": simplified notices providing basic information supplemented by more complete privacy statements (OECD, 2006). Another example is "just in time" notices that aim to deliver messages to an individual at the moment when they are most likely to be of use. Developing privacy icons is another proposal to simplify and improve the communication of information about privacy practices (LIBE, 2013). Increasingly, feedback and awareness tools are being made available to show individuals the possible related consequences of activities that they and others may perform within a particular system. Continued development of new ways of effectively presenting information to individuals can help address the complexity of DDI.

The transparency called for in the "Openness" principle of the OECD Guidelines serves a broader purpose than its direct value to individuals in exercising their access rights; it enables enforcement authorities, privacy advocates, journalists and the general public to better understand and evaluate privacy practices.

Responsible usage and effective enforcement

Focusing more explicitly on promoting responsible usage by organisations could be a useful complement to efforts to improve transparency, access and empowerment. Further efforts are needed to find ways to express boundaries outside of which responsible organisations should not use the fruits of data analytics.

One way to make organisations and individuals aware of the limits to responsible uses – as well as common mistakes and the potentially adverse effects of data analytics – is through education and awareness, which are specifically identified in the revised OECD Privacy Guidelines' call for "complementary measures". Privacy frameworks are articulated in terms of high-level principles that need to be applied in order to ensure effective implementation in practice.

Policy makers and enforcement authorities would need to play a role in helping organisations to identify appropriate substantive limits. Examples can be drawn from guides to credit scoring, policies against the use of genetic information by insurers, and prohibitions on the use of social networking data by employers.

The new provisions in the OECD Privacy Guidelines on implementing accountability respond directly to these new concepts and emerging business practices. Greater emphasis within organisations on internal processes to assign responsibility and to assess (and reassess) risks and controls should improve decisions about responsible usage in those organisations. Some have suggested that one way to treat these issues is through an ethical lens, subjecting data-driven decision making to oversight by experts in data ethics (Mayer-Schönberger and Cukier, 2013), an approach that is gaining momentum (see for example Richards and King, 2014; Johnson and Henderson-Ross, 2012).

Strengthening the enforcement tools of privacy authorities is also needed if these bodies are to play a greater role in monitoring responsible uses. The revised OECD Privacy Guidelines already provide that governments should equip authorities with the resources and technical expertise to exercise their powers effectively; the need for resources and expertise will be all the greater if they are given greater responsibilities to monitor use.

Privacy-enhancing technologies

There are a number of technologies that aim to preserve both privacy and functionality (see Box 5.2). Examples may include privacy-preserving analytics, differential privacy and anonymous credentials. De-identification is often a practical method for obtaining the benefits from data analytics while minimising privacy risks. These risks cannot be completely eliminated, however, and where a sufficient number of different sources of de-identified data are combined, patterns can be revealed that may eventually be traceable back to an individual (EU WP29, 2014a). Some in fact doubt that such tools can withstand progress in re-identification methods, questioning their efficacy as a dependable policy response (PCAST, 2014, p. 39).

Actual risks with regard to re-identification will depend on the context in which the analytics are to be carried out, as well as on the resources and motivation of those who might re-identify. Risk assessment should also consider the likely consequences to the data subject in the event that re-identification occurred. Approaches that combine technical de-identification measures with administrative and legal measures (e.g. enforceable commitments not to re-identify) can help minimise linkability risks (FTC, 2012).

Other types of technical tools can record and describe the life cycle of personal data collected by an organisation (such as provenance) and may assist organisations in managing personal data, and as well as facilitate accountability. For example, advanced data-tagging schemes may be able to attach context and preference information to the data, to help govern future uses (EOP, 2014, p. 56).

Safeguards such as functional separation may also have a useful role to play in ensuring that data used for statistical or other research purposes cannot be used to take decisions with respect to a particular individual (EU WP29, 2013, p. 30). At the same time, functional separation can also support uses related to gaining appropriate insights and knowledge. Policy makers could consider stimulating further research and development in this area, and promote adoption via private-public partnerships, certification schemes and similar initiatives.

Finally, it should be noted that it is possible that the same techniques of data analytics that create discriminatory impacts can likewise be deployed to help individuals and groups identify and assess discriminatory practices, and thereby aid in the enforcement of their rights (EOP, 2014, p. 65).

Privacy risk management

The revised OECD Privacy Guidelines introduce risk management as a key theme for privacy protection, especially in the context of developing privacy management programmes to implement accountability. Risk assessment can consider data sources and quality as well as the sensitivity of the intended uses. In addition to mitigating against the risks of misuse, the assessment can also examine the process by which the data have been analysed; this can help identify where errors or mistakes may have been introduced into the analytical process itself. To be effective, the scope of any privacy risk assessment must be sufficiently broad to take into account the wide range of harms and benefits, yet sufficiently simple to be applied routinely and consistently. It is a challenging task, involving the identification of relevant risks, which may be subjective, and then determining their possible severity and likelihood of impact. As noted above, mitigation measures involving technical tools to de-identify data – particularly when combined with public commitments not to re-identify – may have considerable value, even if they cannot serve as a full guarantee of protection.

Risk-based approaches are not entirely new to privacy frameworks. Risk assessment is implicit, for example, in the OECD Guideline's security safeguards principle[13] and there are close links to privacy impact assessments. Nevertheless, the extent to which a comprehensive risk management approach can strengthen application of the privacy principles is a topic for continued work. As explained in the above section on digital security, successful risk management requires both understanding the risk culture and establishing a risk management framework. The culture of risk management requires a shift from the protection of an asset or an environment from threats, to the optimisation of benefits by recognising that a certain level of risk has to be accepted. How would this translate to privacy protection, where the risks to the individual need to be treated independently of the risks to the organisation?

Risk management also requires one to set the acceptable level of risk, and to treat the risk accordingly on the basis of a full risk assessment. Complex questions remain to be explored further to apply this to privacy protection, such as how to allocate responsibility and how to define the acceptable level of risk. Further, the risk management framework requires establishing a full and ongoing risk management cycle, where awareness, skills,

responsibility and co-operation play key roles, and where risk assessment and treatment are continuous in order to take into account the dynamic nature of the activities and the environment. Finally, the risks to the organisation need to be separated from the risks to the individual. There may be a useful role for third party accreditation in some situations, to validate internal processes for implementing risk-based approaches. Further work is needed to understand how such a framework would best be translated to support the existing privacy protection principles.

5.3. Key findings and policy conclusions

The current and potential benefits of data-driven innovation – building on the expanding availability of data, improved tools to link, process and store them, and algorithms for deriving insights – are amply discussed in other chapters, and have not been addressed here. The focus here has instead been on the importance of addressing the trust issues, the challenges of doing so, and possible ways forward.

Addressing security in the context of data-driven innovation raises a double challenge: first, shifting the culture and mindset of decision makers and other parties involved in DDI, from traditional to risk-based digital security; second, systematically implementing an ongoing risk management process in the overall data value cycle and within each of its parts. The complexity of applying digital security risk management to activities such as data-driven innovation should not be underestimated. It requires a high degree of systematisation and significant management efforts, bundled with operation of the data value cycle itself. Notably however, this is the same risk management approach routinely applied by many businesses in other spheres of their activities to increase their likelihood of success. Revision of the OECD Security Guidelines offers an opportunity to elevate the need for attention to digital security risks to the highest levels in organisations – a key to progress in this area.

The privacy challenges described are not really new. Risks of discrimination are at the heart of privacy laws dating back to the 1970s. As early as 1980, the OECD called for measures to prevent unfair discrimination in its Privacy Guidelines. But the data-intensive developments described in this book do bring the privacy challenges into greater focus.

Several particular points also emerge as showing promise to improve user trust in data-driven innovation. Privacy risk management has been identified as important, but considerable work is still needed to understand how to implement a risk approach for privacy. The experience of the security risk management community may be usefully brought to bear in helping privacy professionals make progress in this area, and this is a topic for future work at the OECD.

Privacy-enhancing technologies continue to offer promise, both in reducing the identifiability of individuals, and in improving the traceability and accountability of policies to protect privacy. Transparency, access and empowerment remain essential to any effective privacy framework, and efforts to improve these dimensions are important. Data access and portability measures can help minimise the information asymmetries and power imbalances that favour data-intensive organisations.

Perhaps the most difficult policy prescription advanced in this chapter is a need for greater effort to articulate substantive boundaries within which responsible uses would be limited. Determining where these boundaries lie – and who should make this determination – will become an increasingly necessary task for organisations making good on the promise of data-driven innovation.

Notes

1 The 2002 OECD *Recommendation concerning Guidelines for the Security of Information Systems and Networks: Towards a Culture of Security* ("Security Guidelines") is being revised at the time of writing.

2 Various alternative models coexist, but the AICIC triad is the most universally recognised.

3 An analogy may be useful here: research shows that biological systems use diversity as a powerful strategy to remain open while successfully responding to the multitudes of evolving threats constantly attacking them. See Forrest, Hofmeyr and Edwards, 2013.

4 For example, those in public and private organisations, who are ultimately responsible for the realisation of economic and social objectives related of data-driven innovation.

5 There are many definitions of risk in various standards. The concept of risk as the effect of consequences on objectives is borrowed from ISO risk standards: ISO 31000 and ISO Guide 73, as well as ISO 27000:2012.

6 Instead of being reduced, risk can also be taken; transferred to someone else; or avoided by not carrying out the activity.

7 This is a simplification: some actors may have a dual role and other categories of actors could also be considered.

8 Data, information and knowledge are seen as different but interrelated concepts. Information is often conveyed through data, while knowledge is typically gained through the assimilation of information. The boundaries between data, information and knowledge may not always be clear, and these concepts are often used as synonyms in media and literature. However, separating the concepts is important to gain a better understanding of data-driven value creation. One can have a lot of data, but not be able to extract value from them when not equipped with the appropriate analytic capacities (see Chapter 3 of this volume). Similarly, one can have a lot of information, but not be able to gain knowledge from it, a phenomenon nowadays better known as "information overload." As observed by Herbert Simon, "a wealth of information creates a poverty of attention" (Shapiro and Varian, 1999).

9 The term is more fully defined and discussed in Chapter 3 of this volume.

10 For example, discrimination in access to fares among passengers on the basis of their place of residence or nationality can infringe EC Regulation No. 1008/2008 of the European Parliament and Council on common rules for the operation of air services in the European Union: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008R1008. More generally, the European Union's Article 29 Data Protection Working Party gives the example of price discrimination based on the computer type used for online purchases as a problematic example of an incompatible use of data (EU WP29, 2013, p 23).

11 For example, the commercial data broker Acxiom has created a website for consumers to gain access to data held about them – www.aboutthedata.com -- in response to the "Reclaim Your Name" initiative by a Commissioner of the Federal Trade Commission (Brill, 2013).

12 See: http://fing.org/?-MesInfos-les-donnees-personnelles-&lang=fr.

13 The EU Article 29 Working Party notes a number of areas in the EU framework (EU Art. 29, 2014c). Risk assessment is required by Korea's certification system (Personal Information Management System).

References

Abrams, Martin (2014), "The origins of personal data and its implications for governance", OECD Expert Roundtable Discussion, Background Paper A – Session I, 21 March 2014 http://informationaccountability.org/wp-content/uploads/Data-Origins-Abrams.pdf.

Brill, Julie (2013), "Reclaim Your Name", 26 June, www.ftc.gov/speeches/brill/130626computersfreedom.pdf (accessed 24 April 2015).

Cavoukian, A. and D. Castro (2014), "Big data and innovation, setting the record straight: De-identification does work", Office of the Information and Privacy Commissioner, Ontario, 16 June, www.privacybydesign.ca/index.php/paper/big-data-innovation-setting-record-straight-de-identification-work.

Council of Europe (2010), "Recommendation on the protection of individuals with regard to the automatic processing of personal data in the context of profiling", CM/REC(2010)13, 23 November, https://wcd.coe.int/ViewDoc.jsp?id=1710949.

Cox, J. (2001), "Can differential prices be fair?", *Journal of Product and Brand Management*, Vol. 10, pp. 264-76, www.emeraldinsight.com/journals.htm?articleid=857764.

Dickson, P. and R. Kalapurakal (1994), "The use and perceived fairness of price-setting rules in the bulk electricity market", *Journal of Economic Psychology*, Vol. 15, pp. 427-48, www.sciencedirect.com/science/article/pii/016748709490023X.

Dwork C. and A. Roth (2014), "The Algorythmic Foundations of Differential Privacy", *Foundations and Trends in Theoretical Computer Science*, Vol. 9, Nos. 2-4, pp. 211-407, http://dx.doi.org/10.1561/0400000042.

EDPS (European Data protection Supervisor) (2014), "Privacy and competitiveness in the age of big data: The interplay between data protection, competition law and consumer protection in the digital economy", March, https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2014/14-03-26_competitition_law_big_data_EN.pdf.

EOP (Executive Office of the President, United States) (2015), "Big data and differential pricing", February, www.whitehouse.gov/sites/default/files/docs/Big_Data_Report_Nonembargo_v2.pdf, accessed 24 April 2015.

EOP (2014), "Big data: Seizing opportunities, preserving values", www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, accessed 24 April 2015.

EU WP29 (Article 29 Data Protection Working Party, European Union) (2014a), "Opinion 05/2014 on Anonymisation Techniques" adopted 10 April 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf.

EU WP29 (2014b), "Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC", adopted 9 April, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

EU WP29 (2014c), "Statement on the role of a risk-based approach in date protection legal frameworks", adopted 30 May 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

EU WP29 (2013), "Opinion 03/2013 on purpose limitation", adopted 2 April, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf.

Field Fisher Waterhouse (2012), "Will access to midata work?", 19 November, http://privacylawblog.ffw.com/2012/will-access-to-midata-work, accessed 24 April 2015.

Forrest, S., S. Hofmeyr and B. Edwards (2013), "The complex science of cyber defense", *Harvard Business Review*, 24 June, http://blogs.hbr.org/2013/06/embrace-the-complexity-of-cybe/.

Frischmann, B.M. (2014), "Human-focused turing tests: A framework for judging nudging and techno-social engineering of human beings", draft paper, 22 September, http://dx.doi.org/10.2139/ssrn.2499760, accessed 23 April 2015.

FTC (Federal Trade Commission, United States) (2014) "Data brokers: A call for transparency and accountability", May, www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014, accessed 24 April 2015.

FTC (2012) "Protecting privacy in an era of rapid change: Recommendations for businesses and policy makers", www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf, accessed 24 April 2015.

Garbarino, E. and O.F. Lee (2003), "Dynamic pricing in Internet retail: Effects on consumer trust", *Psychology and Marketing*, Vol. 20, pp. 495-98, http://onlinelibrary.wiley.com/doi/10.1002/mar.10084/abstract.

Hurwitz, J. (2012), "The making of a (big data) president", *Bloomberg Businessweek*, Management Blog, 14 November, www.businessweek.com/articles/2012-11-14/the-making-of-a-big-data-president.

Intel (2014), "Survey: Distrust and lack of understanding in data privacy fact sheet", www.intel.com/newsroom/kits/bigdata/pdfs/Privacy_Survey_Factsheet.pdf, accessed 24 April 2015.

IWGDPT (International Working Group on Data Protection in Telecommunications) (2014), "Big data and privacy: Privacy principles under pressure in the age of big data analytics", www.datenschutz-berlin.de/attachments/1052/WP_Big_Data_final_clean_675.48.12.pdf, accessed 24 April 2015.

Johnson, D. and J. Henderson-Ross (2012), "The new data values", *Aim*, www.aimia.com/content/dam/aimiawebsite/CaseStudiesWhitepapersResearch/english/WhitepaperUKDataValuesFINAL.pdf, accessed 24 April 2015.

Levine, M.E. (2002), "Price discrimination without market power", Discussion Paper No. 276, 2/2000, Harvard Law School, Cambridge, MA, www.law.harvard.edu/programs/olin_center/papers/pdf/276.pdf.

Lessig, L. (1999), *Code and Other Laws of Cyberspace*, 30 November, Basic Books.

LIBE (Committee on Civil Liberties, Justice and Home Affairs, European Parliament) (2013) "Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data", 22 November, www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2F%2FEP%2F%2FTEXT%2BREPORT%2BA7-2013-0402%2B0%2BDOC%2BXML%2BV0%2F%2FEN&language=EN.

Mayer-Schönberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt.

Merelli, E. and M. Rasetti (2013), "Non locality, topology, formal languages: New global tools to handle large data sets", International Conference on Computational Science, ICCS 2013, *Procedia Computer Science 18,* pp. 90-99, http://dx.doi.org/10.1016/j.procs.2013.05.172.

Mivule, K. (2013), "Utilizing Noise Addition for Data Privacy, an Overview", Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), pp. 65-71, http://arxiv.org/pdf/1309.3958.pdf, accessed 22 April 2015.

Morrone, A., N. Tontoranelli and G. Ranuzzi (2009), "How good is trust? Measuring trust and its role for the progress of societies", *OECD Statistics Working Paper*, Paris.

Narayanan, A. (2013), "Personalized coupons as a vehicle for perfect price discrimination", 25 June, http://33bits.org/2013/06/25/personalized-coupons-price-discrimination, accessed 24 April 2015.

Narayanan, A. and V. Shmatikov (2006), "How To Break Anonymity of the Netflix Prize Dataset", CoRR, abs/cs/0610105, 05 December, http://arxiv.org/abs/cs/0610105.

New York Times (2014), "Protecting student privacy in online learning", 24 September, www.nytimes.com/roomfordebate/2014/09/24/protecting-student-privacy-in-online-learning (accessed 24 April 2015).

Nickerson, D.W. and T. Rogers (2014), "Political campaigns and big data", *Journal of Economic Perspectives*, Vol. 28, No. 2, pp. 51-74, http://dx.doi.org/10.1257/jep.28.2.51.

NSTC (National Science and Technology Council, Executive Office of the President, United States) (2013), "Smart disclosure and consumer decision making: Report of the task force on smart disclosure", 30 May, www.whitehouse.gov/sites/default/files/microsites/ostp/report_of_the_task_force_on_smart_disclosure.pdf.

Odlyzko, A. (2003), "Privacy, economics, and price discrimination on the Internet", Working Paper Series of the University of Minnesota, 27 July, http://ssrn.com/abstract=429762.

OECD (2014a), "Summary of the OECD privacy expert roundtable: Protecting privacy in a data-driven economy: Taking stock of current thinking", OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3xz5zmj2mx-en.

OECD (2014b), *Society at a Glance 2014*, OECD Publishing, Paris.

OECD (2013a), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, 11 July, C(2013)79, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2013b), "Privacy Expert Group Report on the Review of the 1980 OECD Privacy Guidelines", OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3xz5zmj2mx-en.

OECD (2012a), Cybersecurity Policy Making at a Turning Point: Analysing a New Generation of National Cybersecurity Strategies, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k8zq92vdgtl-en.

OECD (2011a), *Society at a Glance 2011: OECD Social Indicators*, OECD Publishing, http://dx.doi.org/10.1787/soc_glance-2011-en.

OECD (2010), *Consumer Policy Toolkit*, Paris, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264079663-en, www.oecd-ilibrary.org/governance/consumer-policy-toolkit_9789264079663-en.

OECD (2006), "Making Privacy Notices Simple: An OECD Report and Recommendations", OECD Publishing, Paris, http://dx.doi.org/10.1787/231428216052.

OECD (2002), "Price Discrimination", *Glossary of Statistical Terms*, OECD, Paris, http://stats.oecd.org/glossary/detail.asp?ID=3283, accessed 24 April 2015.

OECD (1997), Recommendation of the Council concerning Guidelines for Cryptography Policy, C(97)62/FINAL, 27 March, http://webnet.oecd.org/oecdacts/Instruments/ShowInstrumentView.aspx?InstrumentID=115.

OFT (Office of Fair Trading, United Kingdom) (2013) "The economics of online personalised pricing", May, http://webarchive.nationalarchives.gov.uk/20140402142426/http://www.oft.gov.uk/shared_oft/research/oft1488.pdf, accessed 24 April 2015.

Ohm, P. (2009), "The rise and fall of invasive ISP surveillance", *University of Illinois Law Review* 1417.

Online etymology dictionary (2014), "Security", etymonline.com, www.etymonline.com/index.php?term=security, accessed 23 October 2014.

Pariser, Eli (2012), The Filter Bubble: How the New Personalised Web is Changing What We Read and How We Think, Penguin Books, April.

PCAST (President's Council of Advisors on Science and Technology, United States) (2014), "Big data and privacy: A technological perspective", www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf, accessed 23 April 2015.

Pfitzmann, A. and M. Hansen (2010), "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management", v0.34, 10 August, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, accessed 23 April 2015.

Putnam, R., R. Leonardi, and R. Y. Nanetti (1993), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton University Press, Princeton, NJ.

Richards, N.M. and J.H. King (2014), "Big data ethics", *Wake Forest Law Review*, 19 May, http://ssrn.com/abstract=2384174.

Rowling, J. K. (1998), *Harry Potter and the Chamber of Secrets*, Bloomsbury.

Rutenberg, J. (2013), "Data you can believe in", New York Times, 20 June, www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html.

Schermer, B.W. (2011), "The limits of privacy in automated profiling and data mining", *Computer, Law & Security Review*, Vol. 27, p. 45-52.

Sweeney, L. (2013), "Discrimination in online ad delivery", 28 January http://dx.doi.org/10.2139/ssrn.2208240

Turow, J., L. Feldman and K. Meltzer (2005), "Open to exploitation: American shoppers online and offline", Annenberg Public Policy Center of the University of Pennsylvania, http://repository.upenn.edu/cgi/viewcontent.cgi?article=1035&context=asc_papers.

Valentino-Devries, J. (2012), "Websites vary prices, deals based on users' information", *Wall Street Journal*, 24 December, http://online.wsj.com/news/articles/SB10001424127887323777204578189391813881534.

Warden, P. (2011), "Why you can't really anonymize your data", *O'Reilly Strata*, 17 May, http://strata.oreilly.com/2011/05/anonymize-data-limits.html.

World Bank (2011), "General principles for credit reporting", *World Bank Consultative Report*, March, para. 20, http://siteresources.worldbank.org/FINANAICLSECTOR/Resources/GeneralPrinciplesforCreditReporting(final).pdf.

Further reading

Cleveland, H. (1982), "Information As a Resource", *The Futurist*, December, http://hbswk.hbs.edu/pdf/20000905cleveland.pdf.

Cooper, J.C. (2013), "Privacy and antitrust: Underpants gnomes, the First Amendment, and subjectivity", George Mason Law & Economics Research Paper No. 13-39, 21 June, http://ssrn.com/abstract=2283390.

Foundation Internet Nouvelle Génération website, http://fing.org/?-MesInfos-les-donnees-personnelles-&lang=fr, accessed 23 October 2014.

FTC (2007), "Federal trade commission closes google/doubleclick investigation", 20 December, www.ftc.gov/news-events/press-releases/2007/12/federal-trade-commission-closes-googledoubleclick-investigation (accessed 24 April 2015).

Gugeshashvili, G. (2009), "Is goodwill synonymous with reputation", *Juridica International* XVI, p. 126-34, www.juridicainternational.eu/public/pdf/ji_2009_1_126.pdf.

Hillman, R.A. and J.J. Rachlinski (2001), "Standard-form contracting in the electronic age", Working Paper Series of the Cornell Law School, http://ssrn.com/abstract=287819.

Marthews, A. and C. Tucker (2014), "Government surveillance and internet search behavior", 24 March, http://ssrn.com/abstract=2412564 or http://dx.doi.org/10.2139/ssrn.2412564.

Obama (President Barack Obama, United States) (2012), Consumer data privacy in a networked world: A framework for protecting privacy and promoting innovation in the global digital economy", www.whitehouse.gov/sites/default/files/privacy-final.pdf, accessed 23 April 2015.

OECD (2013b), "Supplemental Explanatory Memorandum to the Revised Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data", OECD, Paris, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2013d), *Strengthening Health Information Infrastructure for Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264193505-en

OECD (2013e), *The Internet Economy on the Rise: Progress since the Seoul Declaration*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264201545-en.

OECD (2012c), *Improving the Evidence Base for Information Security and Privacy Policies*, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k4dq3rkb19n-en.

OECD (2011b), *The Evolving Privacy Landscape: 30 Years After the OECD Privacy Guidelines*, OECD Publishing, Paris, http://dx.doi.org/10.1787/5kgf09z90c31-en.

OECD (1997), *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, OECD, Paris, www.oecd.org/internet/ieconomy/guidelinesforcryptographypolicy.htm.

Project VRM (2014), Project VRM website, http://cyber.law.harvard.edu/projectvrm/Main_Page, accessed 23 October 20014.

Shapiro, C. and H.R. Varian (1999), *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston, MA.

Thaler, R. H. and C. R. Sunstein (2009), *Nudge: Improving decisions about health, wealth, and happiness*, 24 February, Penguin Books.

World Economic Forum [WEF] (2014), "Rethinking personal data: A new lens for strengthening trust", www3.weforum.org/docs/WEF_RethinkingPersonalData_ANewLens_Report_2014.pdf, accessed 24 April 2015.

Chapter 6

Skills and employment in a data-driven economy

This chapter discusses the implications of data-driven innovation (DDI) on skills and employment, focusing on two challenges in particular: one, DDI may further increase pressure on the labour market, and especially on middle income jobs, as it enables an increasing number of cognitive and manual tasks to be performed by data- and analytics-empowered applications; and two, the demand for data specialist skills may exceed supply on the labour market. The chapter first shows that DDI could lead to structural change in labour markets, and discusses the implications with regard to skills. It then focuses on data specialist skills and competence, the lack of which could prevent economy-wide adoption of DDI and the (re-)creation of jobs. Finally, the chapter discusses the policy challenges for promoting DDI while smoothing structural adjustments, focusing on challenges in i) addressing wage and income inequalities, and ii) satisfying skills and competence needs.

> *It's in Apple's DNA that technology alone is not enough – it's technology married with liberal arts, married with the humanities, that yields us the results that make our heart sing.* (Steve Jobs during the launch of Apple's iPad 2 in March 2011)

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

The analysis of large volumes of (digital) data, now commonly referred to as "big data", is driving knowledge and value creation across society; fostering new products, processes and markets; and spurring entirely new business models (i.e. data-driven innovation, DDI). Algorithmic trading systems (ATS), for example, analyse massive amounts of market data on a millisecond basis to autonomously identify what to stock and when, and at what price to trade (see Chapter 3 of this volume). ATS, a process unheard of a decade ago, now accounts for more than half of all financial market trading in the United States, and almost a third of all financial market trades in Europe (see Figure 3.12 in Chapter 3).

DDI has the potential to disrupt and transform even traditional sectors such as retail, manufacturing, and agriculture, and thereby to boost economic competitiveness and productivity growth across the economy. Some companies in these sectors are taking advantage of DDI as they are becoming more and more service-like, a trend that some have described using the term "servicification" (Lodefalk, 2010). In manufacturing, for instance, companies are increasingly using sensors mounted on production machines and products, taking advantage of the Internet of Things (IoT). This trend, enabled by machine-to-machine communication (M2M) and analysis of sensor data, has been described by some as the "Industrial Internet" (Bruner, 2013) or "network manufacturing" (Economist Intelligence Unit, 2014). Sensor data are used here to monitor and optimise machine operations at a system-wide level, and for after-sale services, including preventive maintenance operations. DDI is thereby enabling a new generation of highly automated factories, such as the Philips shaver factory in Drachten, the Netherlands, which employs only one-tenth of the workforce it employs in its factory in People's Republic of China (hereafter 'China') that makes the same shavers (see Markoff, 2012).

There is little evidence on the effects of DDI, but the few studies available suggest that firms using DDI raise labour productivity faster than non-users. A study of 330 companies in the United States by Brynjolfsson, Hitt and Kim (2011) estimates that the output and productivity of firms that adopt data-driven decision making are 5% to 6% higher than would be expected from their other investments in, and use of, ICTs. These firms also perform better in terms of asset utilisation, return on equity and market value.

A similar study based on 500 firms in the United Kingdom by Bakhshi, Bravo-Biosca and Mateos-Garcia (2014) finds that businesses that make greater use of online customer and consumer data are 8% to 13% more productive as a result. That study is based on a survey by Bakhshi and Mateos-Garcia (2012), but extended by "matching survey responses about data activities with historical performance measures taken from respondents' company accounts, and by conducting an econometric analysis of the link between business performance and data activity while controlling for other characteristics of the business". The analysis shows that, other things being equal, a one-standard deviation greater use of online data is associated with an 8% higher level of total factor productivity (TFP). Firms in the top quartile of online data use are 13% more productive than those in the bottom quartile. Furthermore, the study shows that "use data analysis" and "reporting of data-driven insights" have the strongest link with productivity growth, "whereas amassing data has little or no effect on its own" (Bakhshi, Bravo-Biosca and Mateos-Garcia, 2014). Another study by Barua et al. (2013) suggests that improving the quality and access to data by 10%, by presenting data more concisely and consistently across platforms and allowing it to be more easily manipulated, would increase labour productivity by 14% on average, but with significant cross-industry variations.

Overall, these studies suggest an approximately 5-10% faster productivity growth compared with non-users. However, it should be stressed that these estimates can hardly be generalised, for a number of reasons. First, as illustrated above, these estimated effects of DDI vary by sector and are subject to complementary factors such as the availability of skills and competences, and the availability and quality (i.e. relevance and timeliness) of the data used. But more importantly, these studies often suffer from selection biases, which make it difficult to disentangle the effects of DDI from other factors at the firm level.[1] More studies are therefore needed to better assess the impact of DDI at that level.

In the current context of weak global recovery, DDI has caught policy makers' attention as a *new source of growth* that can boost the productivity and competitiveness of their economies and industries. However, the disruptive nature of DDI may lead to the "creative destruction" of established businesses and markets, and to a structural shift across the economy, in particular within labour markets. With lingering high unemployment in major advanced economies, however, taking advantage of the process of creative destruction induced by DDI will be particularly challenging, for at least two reasons:

1. DDI may further increase pressure on labour market, in particular on middle income jobs which involve a significant share of tasks that now can be performed by data- and analytics-empowered applications. In particular, DDI enables the automation of an increasing number of cognitive and manual tasks. This includes the use of data analytics for a wider range of intellectually demanding tasks, such as the diagnosis of diseases based on analysis of complex information, including from medical documents. It also involves the use of a new generation of autonomous machines and robots that are no longer restricted to very precisely defined environments, and that can be deployed and redeployed at much faster rates compared to current generation robots.
2. The relatively low availability of the critical data specialist skills and competence required for DDI may prove not just a barrier to the adoption of DDI, but also a missed opportunity for job creation. So far there is little cross-country evidence of a skills shortage or mismatch for DDI. However, some have suggested that the demand for data specialist skills exceeds its supply on the labour market. An Economist Intelligence Unit (2012) survey, for instance, shows that "shortage of skilled people to analyse the data properly" is indicated as the second biggest impediment to make use of data analytics (see also MGI, 2011).[2]

That said, DDI provides huge opportunities for business creation across the data ecosystem for start-ups and small and medium enterprises (SMEs); many of these provide new goods and services that could lead to further job creation opportunities, as discussed in much detail in Chapter 2 of this volume. This chapter, however, does not address these (indirect) job creation opportunities induced by DDI.[3]

6.1. "Creative destruction" in labour markets

As highlighted in Chapter 3 of this volume, the ubiquitous deployment of information and communication technologies (ICTs) driven by Moore's Law[4] has led to a number of developments. These include in particular i) sensors interconnected through machine-to-machine communication (M2M) accelerating the "datafication" of the physical world, ii) cloud computing providing practically anyone with super computing power as an utility, and iii) data analytics empowering decision support and automation for almost

every application area. The confluence of these developments can be expected to reach its tipping point once M2M bypasses human communication, and the Internet will truly become the Internet of Things (IoT – see Box 6.1).[5] This signals a new phase of DDI that today is still in its infancy even in the most advanced economies, and in which software empowered by data analytics could become a major source for labour productivity growth. As Andreessen (2011) wrote, "software is eating the world", and the world will be served in big chunks of data (TNO, 2013). As data-driven software "is eating the world", labour markets may undergo a more profound structural change than what has been observed so far during the digital revolution.

Box 6.1. **The Internet of Things – A game changer**

One of the main reasons for the sudden breakthrough in smart technologies – like driverless cars or the next generation of robots – is the Internet of Things (IoT), which embeds physical objects in information flows. In the case of driverless cars, for instance, it is the road infrastructure, other cars, and last but not least web services such as online maps that "tell" a car essentially what it needs to know. So it is not necessary to equip a car with e.g. a technical image system as powerful as the image processing systems of humans for it to be able to drive on its own, as was previously assumed.

The power of the human image processing system is so huge that it is very difficult to develop a technical alternative of comparable power, despite the growing abilities of these alternatives (Lee, 2015). But such a system is not necessarily required, because cars can now receive huge amounts of just the right data needed to drive autonomously. In this way, cars can "know" even more about their environment than a human driver. For these very reasons a great number of robotic applications formerly thought impossible will become possible soon. It is not that the sensor systems of the robot are exceptionally good; rather, it is that all devices and machines in the manufacturing plant will give the robot the information it needs. This may include products, related robots in the production line, or external suppliers, so that a dynamic optimisation of the overall production process is possible.

Some have stressed that the IoT is also interconnecting and empowering humans with smart applications, leading to the emergence of an intelligent "superorganism" in which the Internet represents the "global digital nervous system" (Radermacher and Beyers, 2011; O'Reilly, 2014). For 2030, it is estimated that 8 billion people and maybe 25 billion active "smart" devices will be interconnected and interwoven by one single huge information network.[1] The result is the constitution of a gigantic, powerful "superorganism"[2], based on never-ending communication streams.[3]

1. In this context, communication can be seen as one of the most powerful intelligence-enhancing processes we know.

2. Communication is what glues the components of that superorganism together. It has a quadratic growth behaviour with respect to the number of components involved, because communication can take place between each pair of members of the superorganism. In a sense, this observation implies positive network effects.

3. The implications are discussed in detail by Kapitza (2005), who looks at the development and size of human civilisation over the past 3 million years. See also Radermacher and Beyers, 2011 and Solte, 2009.

Source: Herlyn et al., 2015.

Starting with the digital revolution and the impact of ICTs

The question of the effects of DDI on employment follows the broader discussion about the employment impact of technology, and of ICT more specifically. That discussion is linked to the fundamental question about "technological unemployment" that Keynes, almost a century ago, described as follows:

> *We are being afflicted with a new disease of which some readers may not yet have heard the name, but of which they will hear a great deal in the years to come – namely, technological unemployment. This means unemployment due to our discovery of means of economising the use of labour outrunning the pace at which we can find new uses for labour. But this is only a temporary phase of maladjustment.* (Keynes, 1930)

The debate on technological unemployment has gained momentum recently in light of current debates over the new potential of automation enabled by data and analytics. Looking at available figures from the Bureau of Labor Statistics on labour productivity and private employment for the United States, scholars such as Brynjolfsson and McAfee (2011) have suggested that the long-run positive relationship between labour productivity growth and employment growth may have been broken since the 1990s. In other words, the labour productivity growth enabled by ICTs in the United States seems not to have led to the creation of further employment.

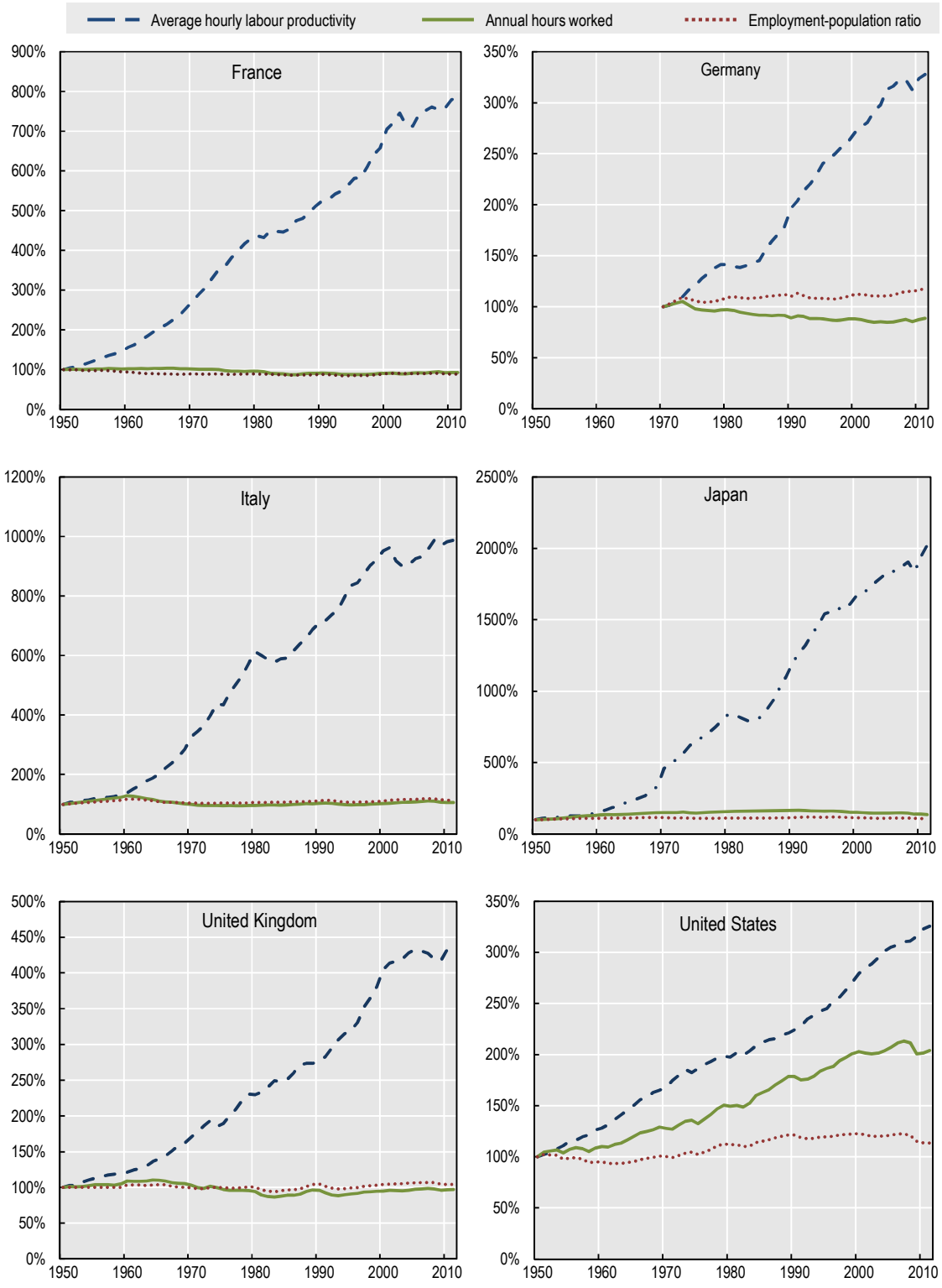
However, such a conclusion is challenged by available cross-country evidence presented in OECD, 2014b, in particular when controlling for the supply of labour. As that report highlights, labour productivity grew at a fast rate in most of the OECD area after the 1960s, while the employment figures in those countries remained stable (Figure 6.**1**). The study therefore concludes that "overall, these long-run trends suggest that compensation mechanisms have been rather effective to maintain employment levels over the last 60 years despite high rates of technological progress" (OECD, 2014b). It should be acknowledged here that this conclusion assumes a "closed" economy where the compensation effects remain within national borders, an assumption challenged by the global and cross-border nature of the data economy. (See the discussion on base erosion and profit shifting, BEPS, in Chapter 2 of this volume.)

Furthermore, many authors – such as Goos, Manning and Salomons (2009) and Autor and Dorn (2013) – have observed a trend of employment polarisation: employment is increasing in both high-skill and low-skill occupations, while stagnating or even declining in middle-skill occupations, with potential negative implications for income equality (OECD, 2014b). "Real wages by skill percentile follow a similar path, suggesting that the increase in employment at the two tails of the skill distribution – high and low skills – has been driven by an increase in demand rather than supply" (OECD, 2014b). According to OECD, 2014b, there are three main explanations for employment polarisation: routinisation, offshoring and international trade. DDI can further leverage routinisation, as it enables and accelerates the automation of some knowledge- and labour-intensive processes.

Furthermore, the increasing use of ICTs across the economy has also driven demand for new types of skills and jobs, most notably linked to ICT specialisation. These are professionals that "have the ability to develop, operate and maintain ICT systems" and for whom "ICTs constitute the main part of their job" (OECD, 2012b). Many of these jobs did not exist before the digital revolution (e.g. software developers), and are rapidly evolving as ICTs progress. In 2013, ICT specialists for the 28 OECD countries considered corresponded to about 14.1 million jobs and to almost 3.5% of total employment, increasing over the decade in all geographic areas (Figure 6.2).

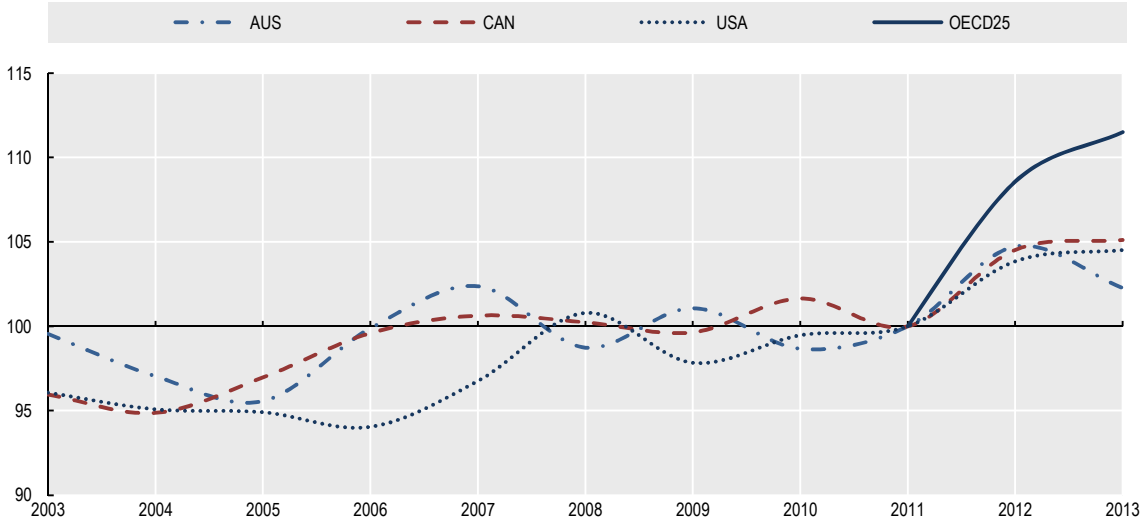
Figure 6.1. **Labour productivity and employment in selected OECD countries (1950-2011)**

1950 (for Germany 1970) = 100%



Source: OECD, 2014b based on Penn World Table, version 8.

Figure 6.2. **Trends in the share of ICT specialists in selected OECD countries, 2003-13**
Index 100 = 2011, share in in total employment



Note: The OECD25 aggregate includes data for all OECD EU member countries plus Iceland, Norway, Switzerland and Turkey. ICT specialists are defined as ISCO 08 codes: 133, 215, 25, 35 and 742. For Australia, Canada and the United States, correspondence tables were used and some adjustments were needed.

Source: OECD based on data from Eurostat, the US Bureau of Census, Statistics Canada, and the Australian Bureau of Statistics labour force surveys.

The growing employment polarisation

Data and analytics have enabled a wide range of "smart" applications that use machine-learning algorithms to "learn" from previous situations, and that can communicate the results of the learning to other machines (see Chapter 3 of this volume).[6] Having analysed similar situations, "smart" applications can infer and predict a present and future situation and therefore can be used for decision automation. They can subsequently perform an increasing number of tasks that are knowledge- and labour-intensive, ranging from search and translation to autonomously operating machines such as cars. This is a new situation that may lead to a deeper transformation through technologies than those seen during the first industrial revolutions, given that more intellectually demanding jobs could be affected. Still, change will be slow, for practical, legal and other reasons.

There is currently a major debate concerning the employment effects resulting from DDI. Many observers see a high risk that "smart" applications will further broaden the employment polarisation highlighted above, at least in the short run. More middle income jobs may be negatively affected – jobs largely held by the segment of the population that "glues" our societies together. Furthermore, DDI will also affect manufacturing by increasing labour productivity,[7] and that could reduce the number of blue collar jobs needed. So while manufacturing could return to OECD economies, at least to some extent, the extent to which this manufacturing on-shoring is likely to generate large numbers of jobs is equally debatable.

Furthermore, this trend may present emerging economies with new challenges, as their role as low-cost assembling points in global value chains may diminish. As a result, there is a risk that one of their historical routes to development – as well as their ability to

leap-frog along it – could be reduced. Brynjolfsson and McAfee (2012a, 2014) and Cowen (2013) observe that after the last big recession in the aftermath of the global financial crisis that started in the year 2007, there was no detectable prospect of job recovery once economic growth took off again. More and more companies consequently announced they intended to replace jobs with machines. A prominent example is provided by the company Foxconn, one of the biggest companies for electronic products, which announced it would replace human workers by 10 000 robots in China (c|net, 2012; Stewart-Smith, 2012; Kan, 2013; Spiegel Online, 2014). Unfortunately, this development will not mean significantly more jobs in the developed world, either. And it has raised concerns among government in emerging economies including in China where Hon Hai Precision Industry is one of the largest private employers (Mozur and Luk, 2012).

The academic debate about technological unemployment will thus most likely intensify, especially since many of the technologies enabling DDI have still not seen large-scale deployment and their economic impacts are therefore still not fully known. In many ways, the world is today at the dawn of machine learning (ML – see Box 3.3 in Chapter 3 of this volume), at a development stage similar to that of the Internet in 1994: few practical commercial examples have reached maturity beyond their large-scale test phase, and much more that is now in the pipelines of research and development (R&D) labs is yet to come. But as applications develop quickly and become more cost-efficient, new generations of autonomous and semi-autonomous machines and systems will be deployed into every part of the economy, bringing with them the potential to displace work in these environments. This could theoretically lead to workerless factories, as the following sections suggest. Even if it causes only temporary friction in the economy, as Keynes once suggested, it is a development policy makers need to consider. Machine learning is as much about the competitiveness of the economy as it is about labour policy.

The following two sections illustrate how DDI-enabled applications will affect i) white collar and ii) blue collar jobs. As autonomous systems are increasingly able to perform intellectually demanding cognitive tasks including, as noted above, diagnosis of diseases based on the analysis of complex information and translation of complex documents and basic spoken language, questions emerge about the extent to which these systems can automate knowledge-intensive tasks that were until recently the object of more highly skilled white collar jobs. Furthermore, DDI enables a new generation of autonomous machines and robots with much more extensive capabilities that can be deployed and redeployed at much faster rates and more cost-effectively compared to current generation robots; these can affect manual labour-intensive blue collar jobs. Some have therefore argued that a long history of achievements of automated processes driven by data processing is reaching a threshold. The question however remains whether the effects on employment will lead to the replacement of jobs by machines, and/or to their "augmentation" or enhancement as better tools become available to the workforce (Davenport, 2014).

Data-driven decision automation affecting white collar jobs

Data- and analytics-enabled smart applications are not used solely by Internet firms. The example of algorithmic trading systems (ATS) in finance, highlighted previously, now accounts for more than half of all financial market trades in the United States.[8] In health care, as another example, medical records with vital signs, magnetic resonance imaging (MRI) and other medical images, can now be analysed with each record representing a pattern that corresponds to diagnoses, therapies and treatments. Machine learning now makes it possible to develop (autonomous) artificial intelligence (AI)

systems such as IBM's Watson that built on considerable parts of the worldwide knowledge base and are even capable of answering questions posed in natural language. Watson, which successfully competed on Jeopardy! against former winners, is now used to support medical diagnosis and therapy – building on more than 600 000 medical reports, 1.5 million patient records and clinical trials, and 2 million pages of medical journal text as of 2013 (IBM, 2013; Upbin, 2013). Such systems could have a profound impact on a number of jobs in the health sector, including high-skilled jobs such as radiologists and oncologists that involve spending a significant share (if not most) of work time identifying anomalies in medical images (Wang and Summers, 2012; Myers, 2011).

Similar systems have been suggested for other data- or information-intensive tasks such as legal advice and even legal decisions. Some of these systems are already used as the basis for legal analytical services provided by companies such as Lex Machina and Huron Legal. These systems can process millions of legal documents and retrieve the most relevant among them in seconds, and far more widely and thoroughly than people can do (Colvin, 2014). Even more striking is that some of these systems can actually predict court decision outcomes.[9] Katz, Bommarito and Blackman (2014), for instance, have developed a system that correctly identifies 70% of the United States Supreme Court's decisions and correctly forecasts 71% of the votes of individual justices across 7 700 cases, and more than 68 000 justice votes. It is difficult to imagine that these smart systems could replace lawyers and judges, although they may in some cases provide better advice "on whether to sue or settle or go to trial before any court and in any type of case" (Colvin, 2014). However, wide adoption of these systems could mean that young legal assistants may no longer be needed to search for relevant legal documents in the discovery phase of litigation, as many have been used to doing until today.

Even for data analysis, which is discussed below as a new key opportunity for employment, advanced analytic tools are now able to automatically fit thousands of statistical models to the available data and automatically generate and test different hypotheses (Davenport, 2014). A statistician relying on manual hypothesis testing can typically create only a few models per week. This has major implications not only for statisticians, but also for researchers (and their assistances). For example, King et al. (2004) presented a system that "automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using a laboratory robot, interprets the results to falsify hypotheses inconsistent with the data, and then repeats the cycle". Scientists are now further exploring the use of data analytics for automated hypothesis generation, and some have proposed analytical frameworks for standardising this scientific approach. Abedi et al. (2012), for example, have developed a hypothesis generation framework (HGF) to identify "crisp semantic associations among entities of interest". Conceptual biology, another example, has emerged as a complement to empirical biology; it is characterised by the use of text mining for automatic hypothesis discovery and testing. This involves "partially automated methods for finding evidence in the literature to support hypothetical relationships" (Bekhuis, 2006). Thanks to these types of methods, insights were possible which otherwise would have been difficult to discover. One example is the discovery of adverse effects to drugs (Gurulingappa et al., 2013; Davis et al. 2013).

New generation of autonomous machines affecting blue collar jobs

Traditionally, robots have been used mostly in manufacturing where their speed, precision, dexterity and ability to work in hazardous conditions are valued. Traditional

robots, however, were fast only in very precisely defined environments; setting up a robotic plant would take months if not years, to precisely plan all the movements of the robots down to the millimetre. Similarly, logistical robots that move the finished components have a precisely choreographed route. The robots might have sensors on board but most of the movements had to be pre-planned and programmed, which did not allow for much flexibility in the production of products. For this reason, the production of consumer electronics is still often done by hand, because the life cycle of consumer electronics and time to market is so short that the robotic factory would not be ready to make the current product by the time the successor should be on the market. This is radically changing because of DDI enabled by the IoT (Box 6.1), where sensor data are feeding machine-learning algorithms that often run via cloud computing services. As a result, machines are becoming more flexible and autonomous and can now perform a wider range of more complex manual work. To understand the employment implications, it is worth recalling the employment impact of the Jacquard loom, the first "programmable" mechanical weaving loom used on a large scale in the textile industry during the early 19th century (Box 6.2).

The potential of autonomous machines is best illustrated with autonomous vehicles such Google's driverless car, which collects data from all the sensors connected to the car (including video cameras and radar systems) and combines it with data from Google Maps and Google Street View (for data on landmarks and traffic signs and lights). If these autonomous vehicles are a success, then autonomous taxis, buses and trucks will be likely candidates for deployment. The effect could be that employment that in the past absorbed unskilled or low-skilled workers will no longer exist. There will still be jobs associated with providing these functions. However, many of them will require higher skills, for example for repairs and programming of robotic functions. Having a skilled labour force is therefore crucial.

Large warehouses have so far also been major employers of workers. In traditional warehouses, the workers walk with pick lists that indicate which items to pick. Modern warehouses use digital technology to direct workers to particular shelves and tells them what items to pick. The worker then scans the barcodes of the items picked and deposited. Workers walk many kilometres each day.[10] Other warehouses use conveyer belts for workers to put products on. The humans are controlled by the computer (see section below on "human computing", in particular Box 6.6). However, in some of the warehouses, the model of working has changed. In these warehouses the shelves are coming to the workers, carried by small driving robots such as those manufactured by Kiva Systems, a company acquired by Amazon after the latter started using Kiva's robots. It creates a different type of warehouse, where the workers stand still and the position of the shelves is dynamic. The location of the goods is continuously optimised, so that the most popular products are on the shelves that need to travel the shortest distance.[11] Pointing, a laser shows the worker what product needs to be picked and where it needs to be deposited. The effect is a supremely efficient warehouse that needs fewer workers to handle the same amount of orders.

Box 6.2. **The Jacquard loom: A driver of industrial revolutions**

In 1801, Joseph Marie Jacquard, a French weaver and merchant, first demonstrated his more highly developed mechanical weaving loom, the Jacquard loom. Mechanical weaving looms had existed before, and had a breakthrough with the invention of the "wheeled shuttle" or "flying shuttle" by the English merchant and inventor John Kay in 1733 (Carlisle, 2004, Kessler, 2004). The key innovation of the Jacquard loom, however, was that it was controlled by an unlimited chain of replaceable punched cards, which enabled the Jacquard loom to be (re-)programmed for the manufacturing of a variety of textiles with different complex patterns. This was key to the Industrial Revolution of the late 18th century, in many respects:

1. The Jacquard loom can be seen as one of the earliest forms of software enabled technology. Punch cards were storage devices on which the woven pattern to be reproduced were encoded. Because the punch cards where replaceable, multiple patterns could be reproduced and if necessary combined to create even more complex patterns. Before the Jacquard loom, only plain (or at best extremely simple) woven patterns could be mass-produced by mechanical weaving looms. More complex patterns were only possible through manual labour.

2. The Jacquard loom had huge implications for higher-skilled textile workers as well. During the Industrial Revolution, the "traditional" mechanical weaving looms led to the automation of processes, rendering many jobs and skills in the "cottage industries" obsolete (*The Economist*, 2014; Dunne, 2014). Large quantities of textiles could be mass-produced at much faster rates than previously with manual workers, and with economies of scale that significantly reduced production costs. There was, however, one area where "traditional" mechanical weaving looms could not compete with skilled manual workers: the production of textiles containing extremely complex woven patterns and pictures. With the introduction of the Jacquard loom, however, even these tasks could then be performed by machines automatically, and at much the same rate and low costs as the production of textiles with plain woven patterns. As a result, even the more highly skilled textile workers, who once were not affected by automation, suddenly were no longer required in the production of complex woven patterns and pictures. As will be discussed below, a similar trend can also be expected to occur in the near future, with a number of middle income jobs being potentially affected by DDI.

3. Finally, the punch card inspired the first generation of computer storage devices, and is therefore considered an important step in the history of computers (Essinger, 2004), and an enabler of the digital (third industrial) revolution. The idea of punch cards inspired for instance the invention of the "mechanical tabulator" in 1889 (US Patent 395 782) by Herman Hollerith, an American statistician at the United States Census Bureau, who went on to become one of the founders of the company that later became known as the International Business Machines Corporation (IBM). Hollerith used this machine to encode data on punched cards more efficiently, thereby boosting data-processing capacities to a whole new level. Hollerith's tabulator enabled the US Census Bureau to complete its 1890 census within just one year, an operation that in the previous 1880 census had taken seven to eight years (Bruno, 2014). This meant a huge cost reduction for the bureau, that needed to employ more than 46 000 census clerks to collect the data – the cost reduction was estimated at the time to be around USD 5 million compared to manual tabulation (Aul, 1972).

C&S, a large supermarket wholesaler in the northeast of the United States, now operates a warehouse that is fully automated from the moment the pallets arrive from the manufacturer and the plastic wrap is removed, to the moment the pallet is put on the truck for transport to the supermarket.[12] Robots move through the shelves and pick the products, which then are handed to autonomously moving robots that bring the products to palletising robots. Not only is the robotic system faster and more efficient with less spillage, but it also allows for the building of higher and better pallets, decreasing the number of trucks needed to ship products, according to the system's manufacturer, Symbotic.[13]

The problems associated with building this kind of warehouse were mostly computational. Such problems include management of the movements of a few hundred robots, such that they do not have to wait for each other and can move at high speed, and that the breakdown of one robot does not break down the system. Another problem is the correct sequencing of picking products, so that a pallet contains products that are all near each other in the supermarket, but heavy products are at the bottom. Each pallet has to be calculated as a 3-dimensional object, not just by itself but also in relation to other pallets destined for the same truck and store. This is a computationally hard problem, known as the knapsack-problem. The solution to the problem came through work done on computer storage problems, where both in desktop computers and in cloud computing the optimal location for data has to be calculated.

In any case, the end result was a warehouse where one pallet assembly robot can palletise 600 cases per hour, compared to 150 for an experienced palletiser and 75 for a starting palletiser, and therefore an important productivity increase. Similar efforts have now also enabled robots to load trucks for package delivery. Today a robotic arm is capable of loading and even unloading the truck using sensors that measure where the other packages are and a three dimensional model that guides the loading of the truck.[14]

Workers are still not easily replaced. The picking of goods in Amazon's warehouse and the loading of trucks in the C&S warehouse continue to be done by humans.[15] But new robots are being introduced into the workplace that could change the situation. An example is the Baxter robot, a new robot that can work together with human workers. It can be programmed by the workers on the floor by moving its arms around, directing its tools and confirming each movement. The robot can then be programmed to perform a task such as packing or unpacking boxes, carrying items to or from a conveyor belt, counting them and inspecting them (Colvin, 2014). And it can easily be reprogrammed if the task needs to be optimised or changed. It is cheaper than comparable robots (USD 22 000) and can be programmed (again, on the job) in a matter of minutes, unlike traditional industrial robots that require days or weeks of highly specific programming by dedicated engineers. Tasks that the robot is currently performing include the boxing of goods and their transfer from one line to another. In such tasks, two robots can replace one human worker, though the robots do require oversight to the level of one supervisor per twenty robots. One could imagine that this robot will replace warehouse workers in jobs such as picking and packaging. Combined with robotic advances in manufacturing, the large deployment of these robots might one day lead to fully automated production processes, from design to delivery.[16]

Implications for skills and competencies

The growing polarisation of employment described above has implications for the skills and competencies needed by the future working force. Studies that have looked in depth at the jobs that could be affected by the emerging automation opportunities also provide insights on the tasks that remain to be performed by humans and the skills and competence needed for these tasks. Furthermore, DDI provides opportunities for labour productivity growth not just through the reduction of jobs, but also through the "augmentation" or enhancement of existing jobs as better tools become available. In that respect, many of the tools described above will augment the intellectual and physical capacities of existing and new workers, opening a new range of possibilities for addressing current and future societal needs. The following sections discuss i) the mix of skills and competencies needed to perform the remaining and new labour-intensive tasks, and ii) the opportunities that DDI offers to augment and enhance the capacities of white and blue collar workers. Whether the future workforce will be adequately equipped with these skills and competencies will finally depend on the capacities of national education systems to support the development of these skills, as discussed further below in this chapter.

Skills and competencies needed for the remaining and emerging jobs

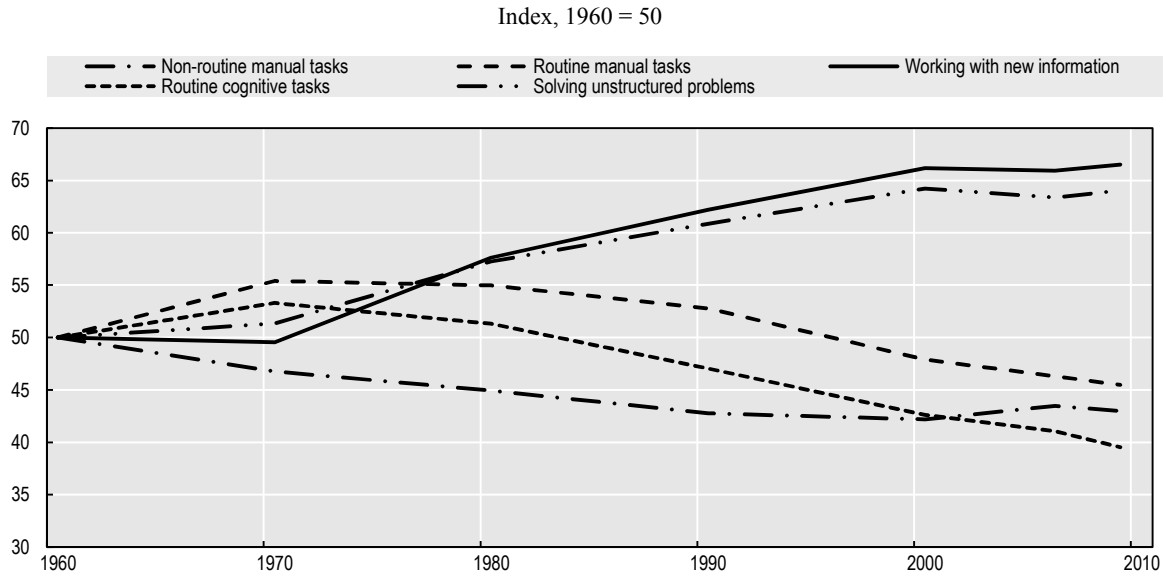
The analysis of jobs affected by automation reveals that a number of tasks will remain to be performed only by humans, and that many of these tasks will become even more relevant for the workforce in the near future. Today, work that consists of following clearly specified directions is increasingly being carried out by computers and by workers in lower-wage countries where the labour costs are still below the costs of machines. The remaining jobs that pay enough to support families will require a deeper level of skills and competence.

Based on the analysis by Autor and Price (2013), Levy and Murnane (2013) note three kinds of tasks on which labour markets will further concentrate (see Figure 6.3):[17]

- *Solving unstructured problems* – including tackling problems that lack rules-based solutions.
- *Working with new information* – including making sense of new data and information for the purpose of problem solving, decision making, or influencing the decisions of others. This includes many of the activities needed for DDI, as will be discussed in the next section.
- *Non-routine manual tasks* – Carrying out physical tasks that cannot be well described via rules because they require optical recognition and fine muscle control (advanced sensomotoric skills) that have still proved difficult for robots to perform.

While solving unstructured problems and working with new information will be particularly important for high-end jobs, carrying out non-routine manual tasks will become increasingly important for low-paying jobs (Levy and Murnane, 2013). This is in line with the observation of employment polarisation presented above – including the findings of Brynjolfsson and McAfee (2011, 2014), who state that a small group of people able to "race with machines" and a large group of people competing for lower-wage job opportunities could result from the DDI that we are witnessing at the moment.

Figure 6.3. **Index of changing work tasks in the United States**



Source: Autor and Price, 2013.

Frey and Osborne (2013, pp. 24-27) provide a similar view on the areas of future employment left for humans, namely jobs that are less likely to be susceptible to computerisation. They argue with reference to other literature that three capabilities in particular remain difficult to automate. Based on the analysis of O*Net data for the United States, they estimate that around 47% of total employment in the United States does not rely on these capabilities, and is therefore seen by the authors as the theoretical maximum share of employment that could be negatively affected by automation. The three capabilities are:

- *Complex perception and manipulation* – Tasks that relate to an unstructured work environment.
- *Creative intelligence* – As creativity involves not only novelty but also values, creative thinking remains out of the realm of computers. Furthermore, creative intelligence is often associated with human intuition as a genuine human capability.
- *Social intelligence* – includes among others the real-time capacity to recognise human emotions and the ability to respond intelligently to such inputs. For computers, this remains a challenging problem.

Elliott (2014) offers a detailed examination of the occupations that are more likely to be affected by automation, based on analysis of the clusters of skills required by these occupations. These clusters include:

- *"Vision movement"* – This includes the combined capacity of i) recognising objects and different features of those objects, including their position in space (vision), and ii) spatial orientation, co-ordination, movement control and body equilibrium. Physical movement is important for jobs in construction, maintenance and production, as well as for jobs in food and personal service. These two large occupational groups represent 30% of current employment.

- *"Language reasoning"* – Including the capacity to deal with natural language: understanding speech, speaking, reading and writing, combined with the capacity to reason, including recognising that a problem exists, applying general rules to solve a problem, and developing new rules or conclusions.

Elliott (2014) suggests that occupations less affected by automation are those that require a high level of at least one of the clusters of skills highlighted above, although the author acknowledges that the progress in automation will soon make even these occupations susceptible to automation in the middle to long term. Elliott (2014) therefore concludes that occupations that involve higher levels of language and reasoning skills are *currently* beyond the capabilities of automation. These include occupations related to education, health care, science, engineering and law.[18] However, as highlighted above, machine learning now makes it possible to develop (autonomous) artificial intelligence (AI) systems such as IBM's Watson that are capable challenging humans even for health care and legal jobs involving higher levels of language and reasoning skills.

Augmenting humans' capacities – Dancing with the machines

As highlighted above, employment opportunities will remain for people with the right mix of skills and competencies. DDI could help augment the intellectual and physical capacities of individuals for these opportunities. Using ever more powerful technical systems enabled by data and analytics as input into the contribution of human work, DDI can further enhance human creativity, social intelligence and sensomotoric skills – and thus combine the experience-based capabilities of analytics with the cognitive capabilities of a highly educated human. This potential has been highlighted by several authors. Brynjolfsson and McAfee (2014), for instance, suggest that we have to learn to race *with* the machines, instead of against the machines, by adding intuition and creativity to the capabilities of new developments driven by big data. Cowen (2013) predicts that the highest performance will be achieved by "freestyle teams", where humans take advantage of their specific know-how and their intuition to best use and connect several systems to get the best results.[19] Levy and Murnane (2013) call this "dancing with robots".

The example of chess is often given to illustrate the power of "dancing with the machines". Currently, even chess software implemented on a smartphone is strong enough to beat most human chess players. In 1997, Deep Blue, a chess-playing computer developed by IBM, defeated the world champion Garry Kasparov. But when it comes to so-called freestyle chess competitions, it is neither humans nor computer systems that win, but the combination of computer systems working with a team of humans. Experience also shows that those humans in "freestyle teams" do not have to be high-level chess players themselves. Their specific know-how is about weaknesses and excellence of all the specific computer systems and how to work with them in a fast and flexible way. They use outputs of systems as an input to other systems, varying and filtering the results. In that way, a network of computer systems is used to derive a viable proposal for the next move to be made in the running chess play. All these computerised decisions are based on data and analytics used by humans with a sufficient level of skills in data and analytics (i.e. data specialist skills – see next section).

The use of data analytics with machines can complement humans' creativity in finding new ways forward and solutions to problems. But as highlighted by Brynjolfsson and McAfee (2014), the combination of humans with machines is also important because at the end of the day it will be humans that will have to take the responsibility for decisions made, as no machine can be held accountable for false decisions. This is

important also because as Taleb (2005, 2010) has demonstrated, "black swans" could lead to false automated outcomes. In other words, many data-driven decisions are based on statistical learning and assume that statistical distribution patterns can always adequately model the reality, including in the future (see Chapter 3 of this volume).

However, many examples show that this assumption does not always hold true; severe economic and social consequences are sometimes the result. Financial crises and the complete failure of modern economics to anticipate the outbreak of the global crisis in 2007-08 can be seen as one example. In medicine, there are cases where the medical data and images used as the basis for diagnostics are not enough to make a correct diagnosis. Professional high-skilled physicians with experience, creativity and intuition and personal interaction with the patient are therefore needed to make the best decisions. Data analytics-empowered machines would provide helpful tools to support their decisions. Those humans teaming with the machines and taking over the accountability and responsibility for their decisions must be highly educated, however. This is because there may be situations where the machine contradicts the opinion of the human decision maker, raising the question whether humans are willing and able to take over the responsibility when overriding a machine's suggested decision (see Box 6.3). Instead of leading us into a future of human-machine collaboration, the future could be a "domination of empiricism" or a "dictatorship of data", where less educated or less concerned decision makers automatically follow the decisions of machines (Mayer-Schönberger and Cukier, 2013).

Box 6.3. **What is new for decision makers with big data?**

In order to understand the implications of big data for decision making, it is important to distinguish big data from conventional information processing. Two major differences deserve highlighting here:

1. Big data is often about providing a kind of "manifest what" by extracting value from a flat and unstructured "datafied universe of information-shreds" with unknown veracity. This helps answer questions on the basis of "alleged insight" via calculated approximations and correlations.
2. Conventional data analytics is about providing a kind of "know what" instead of "manifest what" by extracting "know why" as value. This helps answer questions on the basis of "explicit insight" via revealed causation.

The difference is thus mainly that between correlation and quantitative reasoning compared to causation and qualitative reasoning. The observed trend "from causation to correlation" deserves policy makers' attention where it is leading to "liability aversion". In some cases liability aversion can lead to huge social costs – for instance, when financial risk assessments are based solely on correlations.

Source: Herlyn et al., 2015.

6.2. The growing importance of data specialist skills and employment

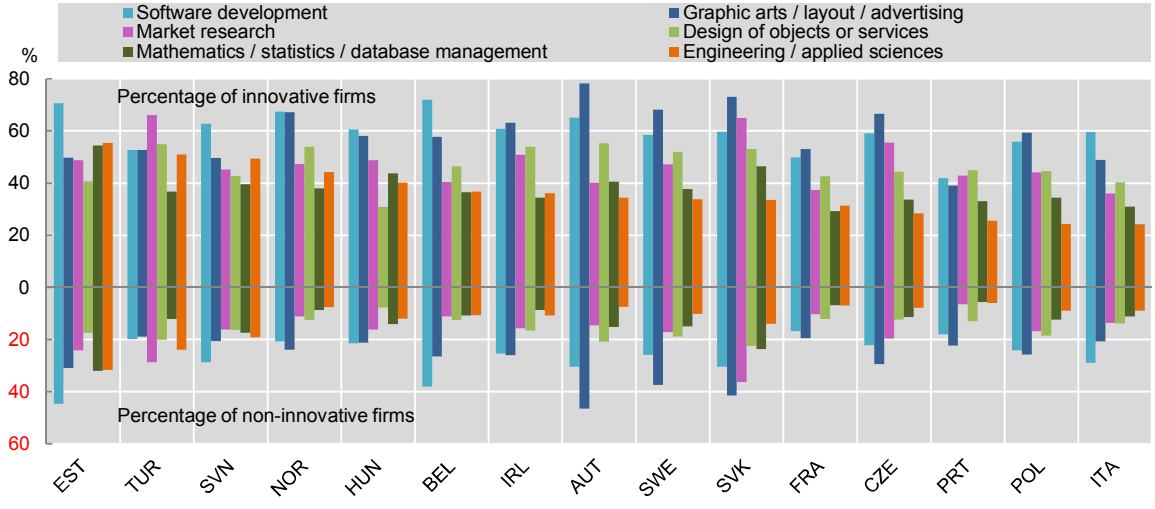
As highlighted in Chapter 2 of this volume, DDI is furthering the creation of new businesses and business models, many of which were unheard of a decade ago, such as data analytic service providers and data brokers and explorers. The results are employment opportunities across the economy. Along with the interest in extracting insights from huge collections of data, the need particularly for people who inherit the skills required for extracting insights from data is growing. This is confirmed by Levy

and Murnane (2013), according to whom "working with new information", including making sense of new data, is one of remaining growing task categories for which labour demand can be expected to increase in the future (see Figure 6.3). Data specialist skills are thus critical for the workforce to be able to "dance with the machines", as discussed above. Furthermore, the right skills in data analytics are essential to understand how to appropriately use data and analytics, and how to deal with the limitations of data-driven decision making highlighted in Chapter 3.

In that respect, data specialist skills are also a key enabler of DDI, as confirmed by business innovation surveys showing that firms using (internal or external) skills related to data and analytics (i.e. mathematics, statistics and database management skills) are more likely to innovate (Figure 6.4).[20] Evidence suggests furthermore that firms with better access to data specialist skills are more likely to gain faster productivity growth through DDI. A recent study by Tambe (2014) was based on an analysis of 175 million LinkedIn user profiles, out of which employees with skills on big data-specific technologies were identified. The study indicates that firms' investment in big data-specific technologies were associated with 3% faster productivity growth, but only for firms that i) already had access to significant data sets and ii) were well connected to labour networks with sufficient expertise in big data-specific technologies. (The estimated output elasticity of 3% resulted after controlling for firms' adoption of data-driven decision making.) This highlights the complementarity effects among data, analytics and skills, the understanding of which merits further study.

Figure 6.4. **Firms using innovation-relevant skills, 2008-10**

As a percentage of innovative and non-innovative firms



Source: OECD Science, Technology and Industry Scoreboard 2013, http://dx.doi.org/10.1787/888932890770.

However, some evidence suggests that the demand for data specialist skills already exceeds the supply on the labour market. The Economist Intelligence Unit (2012) survey shows that "shortage of skilled people to analyse the data properly" is indicated as the second biggest impediment to making use of data analytics. For consumer goods and retail firms it is the single biggest barrier, cited by two-thirds of respondents from those sectors. Some other studies have concluded that there are considerable mismatches between the supply of and demand for data specialist skills. MGI (2011), for instance,

estimates that the demand for deep analytical positions in the United States could exceed supply by 140 000 to 190 000 positions by 2018. This does not include the need for an additional 1.5 million managers and analysts who can use big data knowledgeably. However, further evidence to confirm this trend across countries is needed.

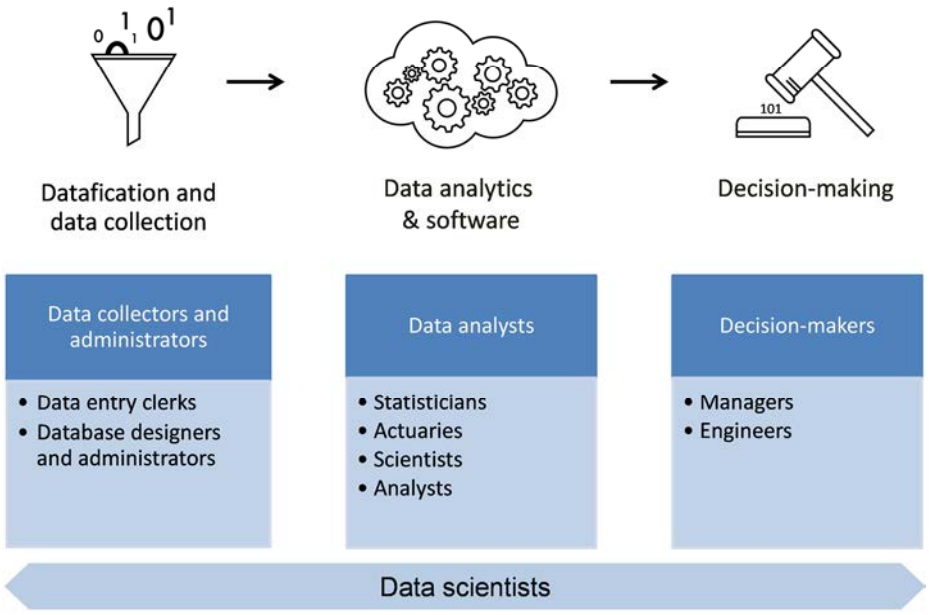
After defining data specialist skills, this section looks at the availability of these skills across the economy and reviews trends in their development over time, including wages and vacancies. Where possible, the section compares these developments with broader economy-wide trends in employment, focusing on ICT specialist employment. Following the methodology applied in OECD, 2013a, insights are provided on the data intensity and diffusion of DDI across the economy from the perspective of skills and employment, and offers a picture of the relative demand for data specialist skills across sectors, highlighting those that intensively employ data specialists. By measuring the share of each sector's workforce related to data, the section provides an approximate figure for the use of data (and analytics) across the economy, a figure that until now could not be provided through official statistics for reasons discussed in Box 2.1 in Chapter 2 of this volume.

Defining data specialist skills and employment

There is currently no commonly adopted definition of data specialist skills. To a large extent this is due to the fact that these skills have not received much attention in literature compared for example to ICT (specialist) skills – which in many respect cover, although not perfectly, what are termed in this chapter data specialist skills. Furthermore, DDI is a relatively new phenomenon that has only recently caught the attention of decision and policy makers, and it is therefore only recently that data specialist skills have risen to the top of the agenda of different stakeholders. And, last but not least, DDI is not only a new phenomenon but also a rapidly evolving one, which means that many data specialist jobs are rapidly evolving as well. Whereas some of these jobs are already established in labour markets, many – such as "data scientists" – are rather new professions that just recently emerged in light of a convergence of disciplines, including computer science and statistics but also natural and social sciences as well as business management, marketing, and finance.

All these data specialist professions have one thing in common: *working with data constitutes a main part of their job*. Different data specialist occupations can be identified using the data value cycle introduced in Chapter 1 as a framework (Figure 6.5). These include in particular occupations that mainly i) collect and/or manage data, such as data entry clerks and database designers and administrators, and ii) analyse data through analytics, including in particular statisticians and actuaries; scientists such as astronomers, epidemiologists and economists; and analysts such as those in finance, market research and intelligence. But it also includes related associate professionals and clerks such as statistical assistants. To a limited extent, data specialist professions also include iii) data-driven decision makers such as managers and engineers, for whom however working with data rarely constitutes a major portion of their jobs.

Figure 6.5. **Main phases of the data value cycle with their key types of data specialist occupations**



Given this definition of data specialists, measuring related occupations based on official statistics remains challenging for several reasons. First, the rapidly evolving nature of data specialist skills has led to the emergence of new professions such as "data scientists" that are not properly captured by official statistics (Royster, 2013). Also, some occupations such as economists may often, but not always, require working with data as a main part of the job. And finally, the official national statistics provided are often not granular enough to really capture data specialist occupations, and this becomes even more of an issue when comparing available statistics across countries. Box 6.4 therefore proposes an operational cross-country definition of data specialist skills that, for comparability and measurement reasons, excludes a number of occupations that would otherwise be captured by the framework presented in Figure 6.5.

Box 6.4. **Data specialists: Towards an operational cross-country definition**

Following the OECD definition of ICT specialist, data specialists are defined for the purpose of this report as those occupations for which *working with data constitutes a main part of the job*. In an attempt to provide comparable measures across OECD countries, data specialists have been defined according to the 2008 International Standard Classification of Occupations (ISCO-08) to include the following two occupations at three-digit level:

- 212 – Mathematicians, actuaries and statisticians
- 252 – Database and network professionals.

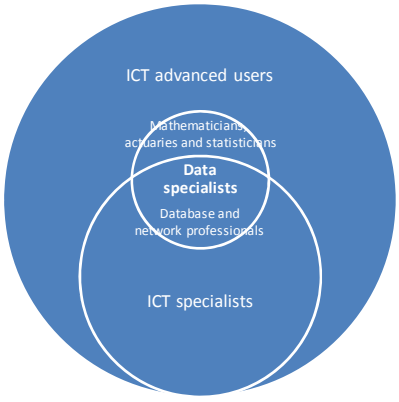
Box 6.4. **Data specialists: Towards an operational cross-country definition (cont.)**

It should be noted that some occupations considered as data specialist are missing. These include in particular "data entry clerks" (4132), which is only available within a larger set of occupations (at 3-digit level) that additionally includes non-data specialists.[1] The same is true for "statistical, mathematical and related associate professionals" (3314), which are also a subset of a much larger non-data-specialist group of occupations.[2] Furthermore, "physicists and astronomers" (2111) are often considered data-intensive occupations, but the 3-digit occupation "physical and earth science professionals" to which they belong not only includes meteorologists (2112) which can be considered data-intensive, but also chemists (2113). Epidemiologists are not explicitly captured by ISCO-08, but are part of a larger group including "biologists, botanists, zoologists and related professionals" (2131). Finally analysts often also include a large number of non-data specialist occupations such as "advertising and marketing professionals" and "public relations professionals", some of which rarely require working with data as a main part of the job.[3]

The proposed definition, including 212 and 252, is therefore considered to best strike the balance in capturing the employment activities related to the use of data across countries, and is seen as a *narrow definition*. With the help of correspondence tables, a comparable list of occupations has been compiled using regional classifications to measure data specialist employment in Australia, Canada and the United States (see Annex Tables 6.A3, and 6.A4). A much *broader definition*, including those occupations highlighted above, deserves further study.

Finally, it is interesting to note that the definition of data specialist proposed here is not a perfect subset of the OECD definition of ICT specialists (which includes database and network professionals), but the definition of data specialist further includes "advanced ICT users" (mathematicians, statisticians and related professionals) that were part of a broader definition of ICT skill occupations presented in OECD, 2005 (see Figure 6.6)

Figure 6.6. **Data and ICT specialists in context**



1. These include "typists and word processing operators" (4131). It should be noted that the omission of these occupations from an operational cross-country definition should not provide any significant challenge, given their declining share in the economy.

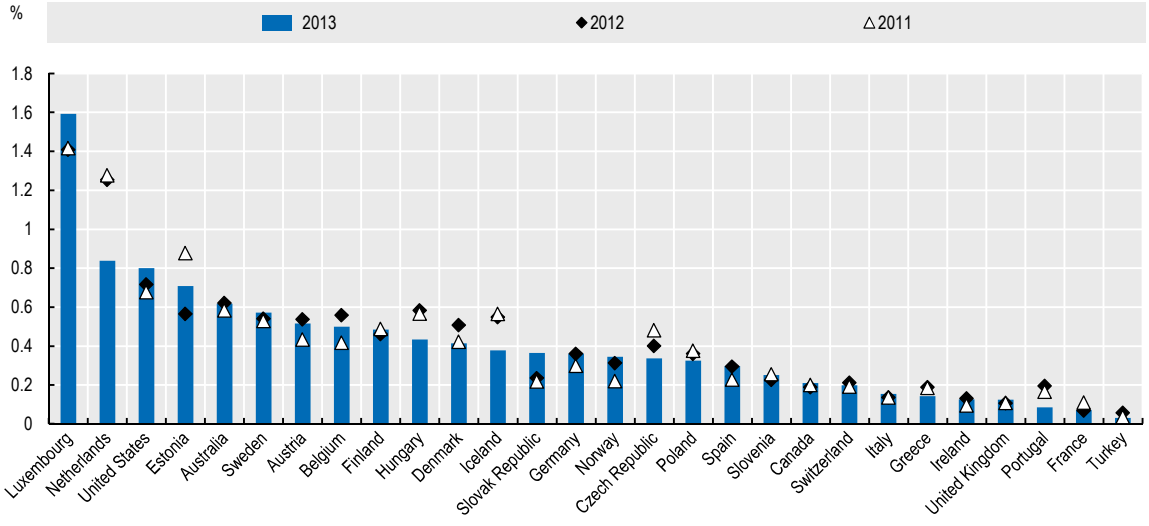
2. The group includes "securities and finance dealers and brokers" (3311), "credit and loans officers" (3312), "accounting associate professionals" (3313), and "valuers and loss assessors" (3314).

3. They include "accountants, financial and investment advisers" and "financial analysts" (241), which can be considered more data-intensive, but also "technical and medical sales professionals" and "information and communications technology sales professionals" (243), which are rarely considered data-intensive.

Estimates based on the definition proposed in Box 6.4 suggest that data specialists in 2013 accounted for over 0.6% of total employment in countries such as the Netherlands, the United States, Australia and Estonia, while in Luxembourg the share of data specialists almost reached 1.6% of total employment (Figure 6.7). In countries such as Portugal, France and Turkey, the share of data specialists is far below 0.1%. In most economies, the share has increased significantly over the past years, suggesting not only that demand for data specialists has increased faster than demand for other types of jobs, but also that these economies have become more data-intensive over time. Employment figures for Canada and the United States show that the share of data specialists in total employment has rapidly increased since 1999, even faster than the share of ICT specialists (Figure 6.8; see Figure 6.9 for Canada). For comparison, ICT specialists in the 28 OECD countries for which data are available corresponded to almost 3.5% of total employment in 2013 (about 14.1 million jobs). There are however some notable exceptions to the growing share of data specialists, namely in the Netherlands, Hungary, Iceland, Denmark, Czech Republic, Poland, Greece, and Portugal.[21]

Figure 6.7. **Data specialists in selected OECD countries, 2011-13**

As share of total employment

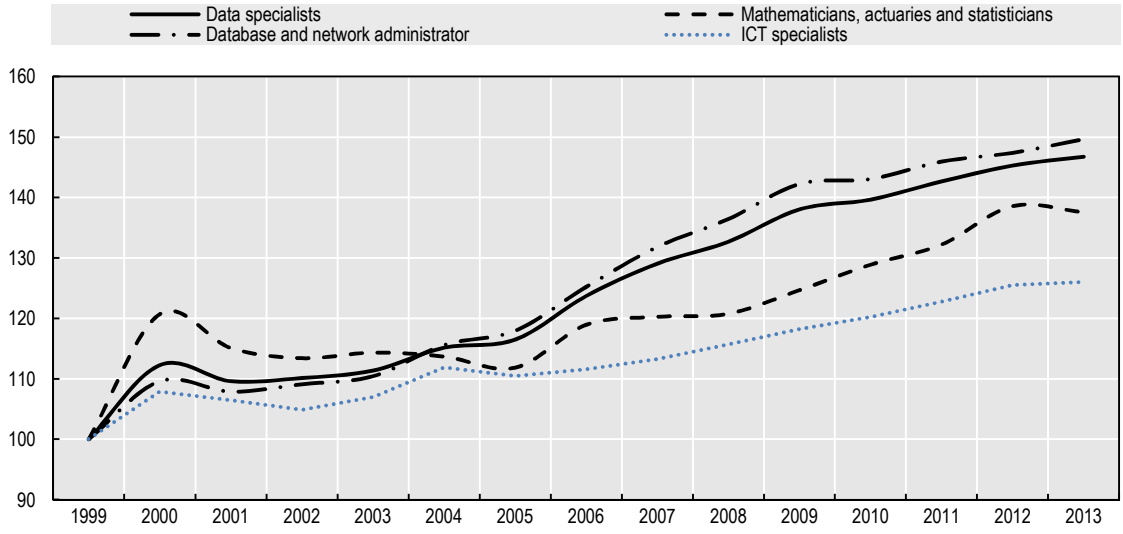


Note: Data for Ireland and the United Kingdom only include ISCO-08 code 212 "mathematicians, actuaries and statisticians" as data for code 252, "database and network professionals", are not available. Data for Canada include the equivalent of ISCO-08 codes: 212 and 252. Data for the United States are overestimated since parts of other ISCO-08 codes (3514 and 2519) are included.

Source: Based on data from Eurostat, Statistics Canada, Australian Bureau of Statistics Labour Force Surveys and US Current Population Survey, March Supplement, February 2015.

Figure 6.8. **Trends in the share of data specialists in the United States, 1999-2013**

Index, 100 = 1999, share in total employment

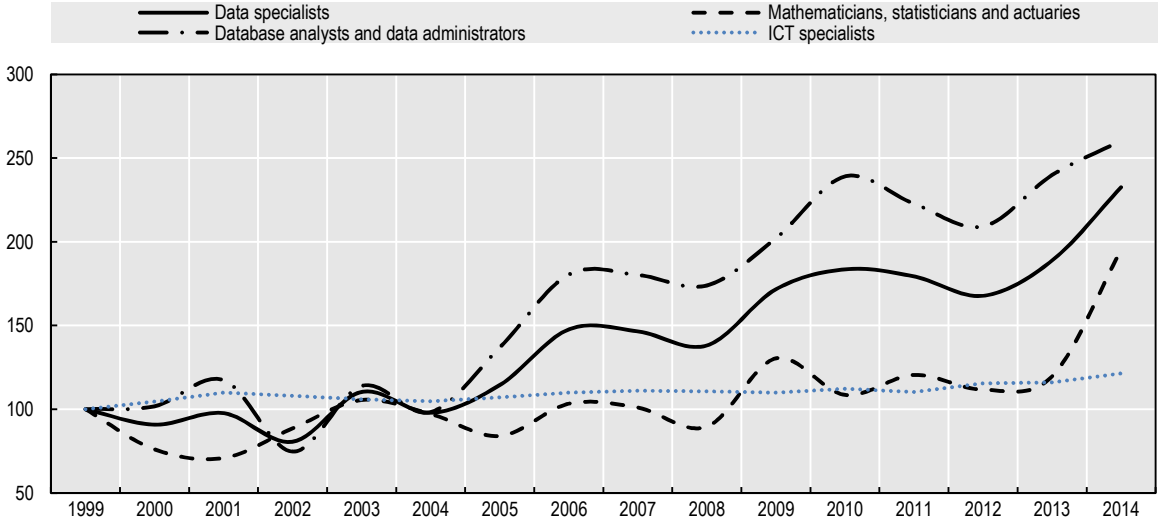


Note: "Data specialists" does not correspond here to the ISCO definition presented in Box 6.4. In order to be consistent across years, the definition has been slightly modified and does not include "information security analysts" (SOC 2010 code 15-1122), "computer network architects" (15-1143) or "computer occupations, nec" (15-1199).

Source: Bureau of Labor Statistics, Occupational Employment Statistics (OES), www.bls.gov/oes/home.htm, November 2014.

Figure 6.9. **Trends in the share of data specialists in total employment in Canada, 1999-2014**

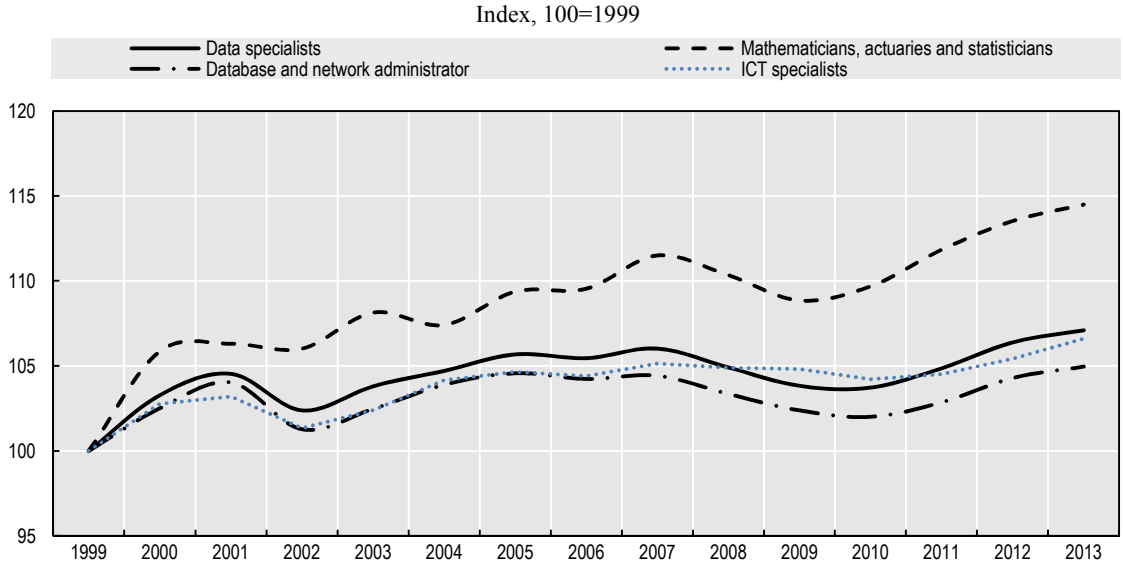
Index, 100 = 1999



Note: "Data specialists" does not correspond here to the ISCO definition presented in Box 6.4. In order to be consistent across years, the definition has been slightly modified, and only includes ISCO 08 code 212, "mathematicians, actuaries and statisticians", and code 2521, "database designers and administrators" (equivalent to NOCS 2011 code 2172, "database analysts and data administrators").

Source: Statistics Canada, labour force survey, February 2015.

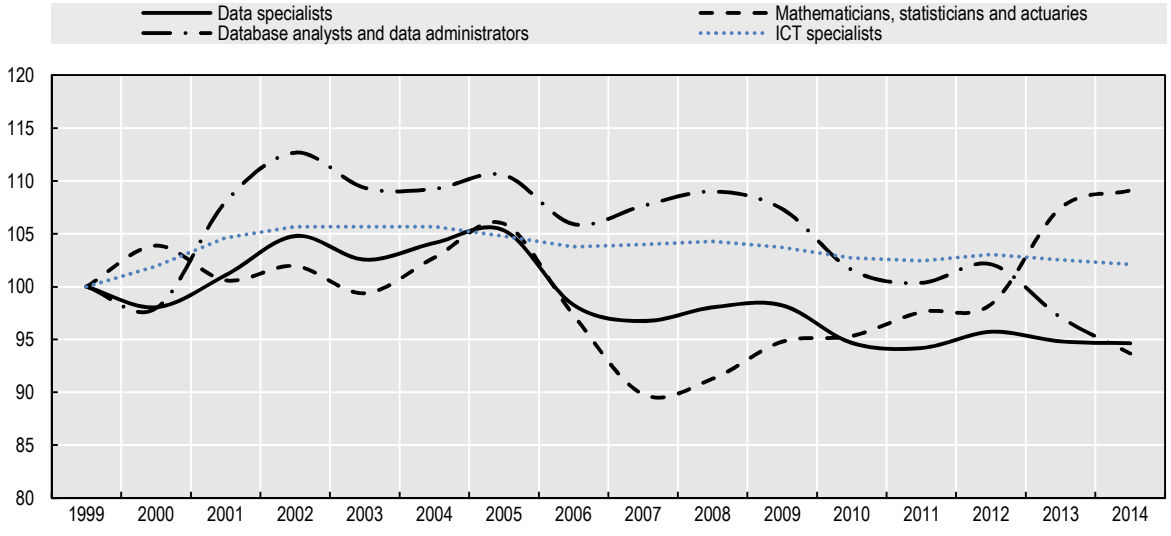
Figure 6.10. **Trends in relative average wage of data specialists in the United States, 1999-2013**



Note: "Data specialists" does not correspond here to the ISCO definition presented in Box 6.4. In order to be consistent across years, the definition has been slightly modified and does not include "information security analysts" (SOC 2010 code 15-1122), "computer network architects" (15-1143) or "computer occupations, nec" (15-1199).

Source: Bureau of Labor Statistics, Occupational Employment Statistics (OES), www.bls.gov/oes/home.htm, November 2014.

Figure 6.11. **Trends in relative average wage of data specialists in Canada, 1998/99-2013/14**



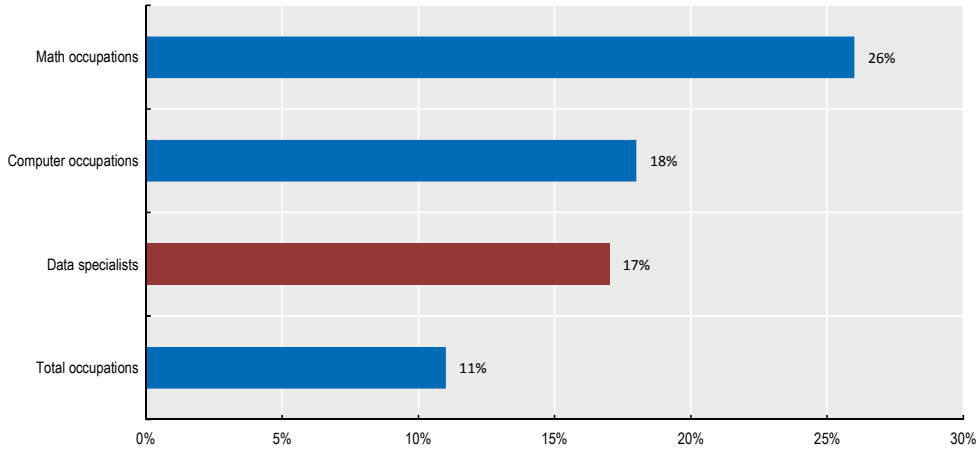
Note: Wage data are a two-year moving average based on average weekly earnings from 1998/99 to 2013/2014. "Data specialists" does not correspond here to the ISCO definition presented in Box 6.4. In order to be consistent across years, the definition has been slightly modified, and only includes ISCO 08 code 212, "mathematicians, actuaries and statisticians", and code 2521, "database designers and administrators" (equivalent to NOCS 2011 code 2172, "database analysts and data administrators").

Source: Statistics Canada, labour force survey, February 2015.

It is worth noting that the ratio of average annual wages of data specialists to that of all occupations has remained relatively stable in the United States in the last decade (Figure 6.10, for Canada see Figure 6.11), suggesting that demand for data specialists was satisfied through labour markets in general. Based on estimates of the US Bureau of Labor Statistics, demand for data specialist jobs are however expected to grow to 17% in the United States between 2012 and 2022 (Figure 6.12). This is six percentage points faster than the estimated total employment growth for that same period. Statisticians, actuaries and mathematicians are expected to have the fastest growth between 2012 and 2022 (26%). This is consistent with the observation that statisticians, actuaries and mathematicians are expected to be in higher demand as DDI becomes more important for businesses. These occupations have also seen the fastest growth in relative wages since 1999, compared to data specialists and ICT specialist for which relative wages have grown more modestly (Figure 6.10). But it is also observable that the share of statisticians, actuaries and mathematicians has been decreasing since 2012, suggesting – along with the further growing relative wages for that group – that the United States could be facing a shortage.

Figure 6.12. **Data specialist jobs outlook in the United States, 2012-22**

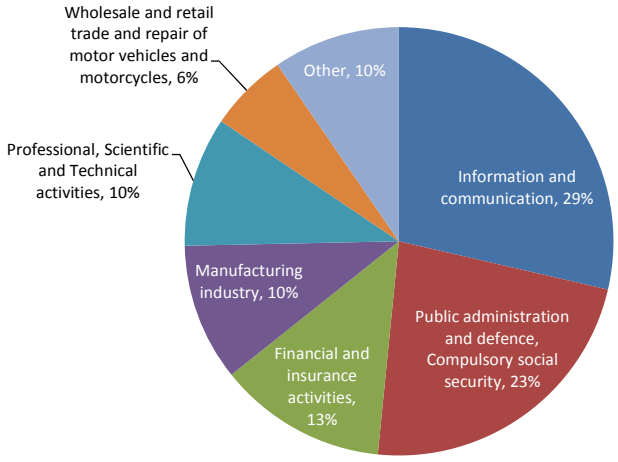
Percentage change in employment, projected 2012-22



Source: Based on the US Bureau of Labor Statistics, Employment Projections programme, December 2014.

Data specialists are more likely to be in highest demand in those economies where data-intensive industries are more prevalent, such as in Luxembourg where the financial sector is a major industry. The most data-intensive industries employing the highest share of data specialists are still the ICT service industries (see Figure 6.13),[22] and in particular i) IT and other information service industries, but also ii) insurance and finance, iii) science and research and development, iv) advertising and market research, as well as v) the public sector including extraterritorial organisations and bodies (such as the OECD, UN and other international organisations). A similar concentration can be observed for the United States. These findings are in line with recent studies suggesting that ICT firms are still leading in the use of advanced data analytics; according to Tambe (2014), only 30% of Hadoop investments come from non-ICT sectors, including in particular finance, transportation, utilities, retail, health care, pharmaceuticals and biotechnology. It is interesting to note that most of the data-intensive sectors also tend to have a high ICT intensity (ICT expenditure as a share of output).[23]

Figure 6.13. **Distribution of data specialists per industry in selected OECD countries, 2013**



Note: Industries are based on ISIC rev. 4 (2-digit codes). OECD 25 includes all OECD EU Member Countries plus Iceland, Norway, Switzerland and Turkey.

Source: European Union Labour Force Survey, November 2014.

Data specialists in transformation

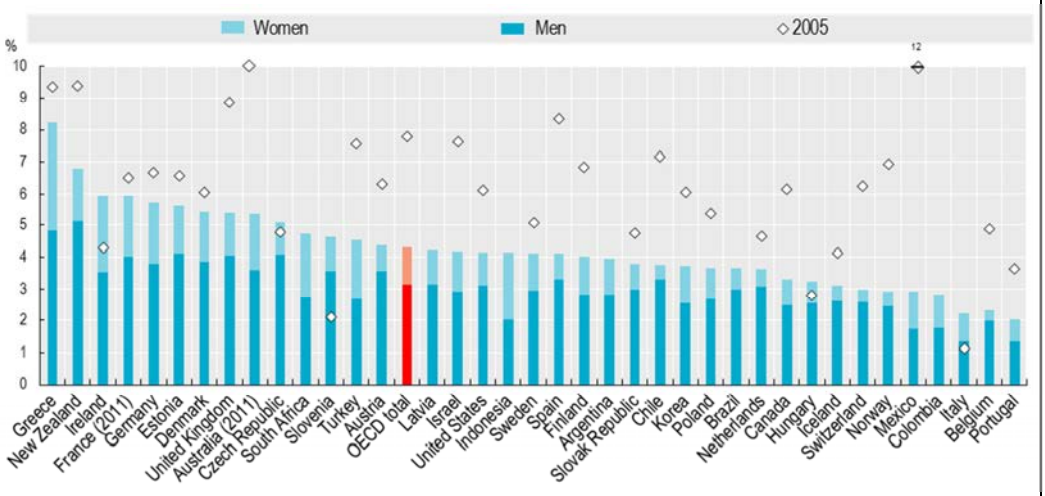
There are a number of sociological and technological trends that are transforming data specialist occupations. One important sociological trend is the increasing participation of women in data specialist occupations, which has not only contributed to a higher diversity, but also led to a higher supply of skills in the labour market (see Box 6.5). Another important (technological) trend is the increasing role of ICTs in data specialist occupations, and in particular the impact of the convergence of disciplines such as mathematics and statistics, but also that of natural and social sciences – such as physics and economics – with computer science. This convergence has led to a new class of data specialist, the "data scientist", a term that is still vague but used by several authors to describe the emergence of a "new" discipline, job category or career path that has grown in importance and prominence with big data. The major data specialist occupations are discussed in more detail in the following sections, beginning with traditional occupations that include data entry clerks and database designers and administrators, and then actuaries, mathematicians and statisticians. The section ends with a discussion of the term "data scientist" itself.

Box 6.5. **The growing importance of women in data specialist occupations**

Women still participate significantly less in the data specialist occupations than men. One out of five data specialists is a woman. However, while their share in overall employment remained stable (0.05%) between 2011 and 2013, the proportion of women increased from 16% in 2011 to 20% in 2013. When looking at the occupations included in the definition of data specialists, it is interesting to note that around 40% of "mathematicians, actuaries and statisticians" are women.[1] The picture is somewhat different for the "database and network professionals", where only 14% are women.

The still relatively low share of women in data specialist occupations is reflected in the low participation of women in tertiary-level graduate programmes related to mathematics, statistics and computer science. In 2012, for instance, the share of data-related graduates among all tertiary graduates in the OECD area was 4.3%, of which 27% were women (Figure 6.14).

Figure 6.14. **Data-related tertiary graduates, by gender, 2005 and 2012**
As a percentage of all tertiary graduates



Note: "Data-related tertiary graduates" has been defined for this figure as persons who have attained a degree in the field of Computer Science (ISC 48) and Mathematics and Statistics (ISC 46) based on the International Classification of Education (ISCED-97), levels 5A and 5B. Data for Luxembourg and Japan are not available, nor are data for Australia or Israel with respect to ISCED 5B data.

Source: OECD Education Database and OECD (2014f), *Education at a Glance 2014: OECD Indicators*, OECD Publishing, Paris, www.oecd.org/edu/eag.htm, accessed 28 June 2015.

Evidence suggests that female data specialists tend to concentrate in sectors such as "public administration and defence"; "compulsory social security" represents almost 20% of all female data specialists, followed by "financial and insurance activities" (17%). "Computer programming, consultancy and related activities" and "information services activities" industries also attract a good number of women (16%). By comparison, male data specialists are concentrated in "computer programming, consultancy and related activities" (over 23%), followed by "financial and insurance activities" (12%). "Public administration activities" comes third, with near 10% of all male data scientists.

1. In 2014, half of the "mathematicians, actuaries and statisticians" were women in Australia. In Canada, the percentage is 44% for the same year.

Traditional data specialist occupations

The main responsibilities of data specialists are the collection, management, processing and analysis of data sets to discover new insights – for example, through statistical modelling – to enable better decisions and predictions about the future. Traditionally these functions have often been undertaken through a division of labour among different occupations along the data value cycle presented in Figure 6.5.

Data entry clerks often stand at the beginning of the data value cycle, given that their main tasks are to collect and enter data into "electronic equipment, computerised databases, spreadsheets or other data repositories using a keyboard, mouse, or optical scanner, speech recognition software or other data entry tools" (ILO, 2009). Data entry clerks are especially needed to transform and store unstructured data (i.e. data that have no predefined data model or structure, such as text and multimedia files) in structured digital data formats. To understand the historical importance of data entry clerks, it is necessary to recall that unstructured data are estimated to still account for between 50% and as much as 85% of all data stored in organisations (see Chapter 3 of this volume).[24] And many data-intensive operations, such as a census, were and still are not possible without data entry clerks (see the discussion of Hollerith's tabulator in Box 6.2).

However, with rising computing capacities, data analytics is increasingly able to automatically extract some structures embedded in unstructured data, including multimedia content (Chapter 3).[25] As a result, data entry clerks have become less and less important for most organisations, a trend reflected in the ever decreasing share of data entry clerks observed in the United States since 1999. In addition, the increasing deployment of sensors with the Internet of Things has significantly increased the potential to automate data collection and storage to such an extent that few domains will remain where data entry clerks will be needed in the near future, including for the collection of census data (see Reimsbach-Kounatze, 2015). In that respect, data entry clerks can be seen as one of the job categories that DDI has successfully rendered less important; ironically, these jobs were initially one of the key enablers of DDI. As will be highlighted further below, this ironic twist is not limited to data entry clerks but also includes more intellectually demanding data specialist jobs, such as even statisticians.

Although the decreasing share of data entry clerks can be seen as an indicator for the increasing capacity of machines to do their jobs, their full replacement by machines is still some way off. As highlighted in the previous section, the capacity of "language reasoning" (Elliott, 2014) remains the competitive advantage of humans over machines, and this capacity has been one of the key reasons why data entry clerks remain employed today. Language reasoning includes the capacity to recognise meanings, which is essential – for example, when extracting related but ambiguous information where different ways of representing the same entity prevents computers from making semantic linkages. The problem is particularly relevant when different data sets need to be linked. The demand for these types of data-related tasks is reflected in the increasing number of crowdsourcing activities that some have referred to as "human computing", and that are offered through services such as Amazon Mechanical Turk (MTurk) since 2005 (see Box 6.6). These services involve small tasks for which human intelligence is required and no cost-efficient algorithm exists; examples include data cleaning and verification, including the classification of data entities and the identification of duplicates.

Box 6.6. **Crowdsourcing of human intelligence tasks: "Human computing" and "micro tasking"**

While computing and automation technologies are steadily improving, there are still many tasks that human beings can do much more effectively than computers, such as identifying objects in a photo or video, performing data de-duplication or transcribing audio recording. To perform these often one-time tasks, firms tend to hire temporary workers. Crowdsourcing, a workforce for human intelligence tasks (HITs), is increasingly used as alternative to solve this problem while providing firms with even more flexibility and scalability when outsourcing these labour-intensive tasks that computers cannot perform. This process is often referred to as "human computing" to illustrate the reverse role between humans and computers, where computers "use" humans to solve problems that computers cannot perform. But the term also refers to the more traditional term "computer", referring to a human that "is supposed to be following fixed rules; he has no authority to deviate from them in any detail" (Turing, 1950).

Amazon is still the most prominent large-scale provider of "human computing" services over the Internet, since it launched its crowdsourcing marketplace for digital work called Amazon Mechanical Turk (MTurk) in 2005. Requesters advertise small projects that cannot be fully carried out by computers on the online platform. Worker called "turkers" can then complete those one-time tasks for mostly a very small amount of money, usually ranging from as little as USD 0.01 for a quick task up to rarely more than USD 100 for more complex jobs. Currently, there are around 500 000 workers from 190 countries registered at Amazon MTurk. Especially for people living in developing countries, MTurk and similar services have been highlighted as a job opportunity to overcome poverty, although the requirements of an Internet connection and English language skills still restrict this potential. Samasource, a nonprofit organisation based in the United States whose mission is "to use work, not aid, for economic development", provides data-related services to large companies in the United States and Europe. It divides the work up into small pieces (called "microwork") and then sends it for completion to delivery centres in developing regions including Haiti, India, Kenya and Uganda (Gino and Staats, 2012).

While they represent job opportunities for some, MTurk and similar services such as Samasource have been criticised, analogous to the socially unacceptable working conditions in the textile industry in the 19th century, for being a "digital sweatshop", given that these services "[circumvent] a range of labor laws and practices, found in most developed countries, that govern worker protections, minimum wage, health and retirement benefits, child labor" (Zittrain, 2009, cited in MIT Technology Review, 2010). Authors such as Uddin (2012) and Cushing (2013) and studies by Horton and Chilton (2010) have stressed that "microworkers" typically work at a by far below average hourly wage (estimated to be less than USD 1.50). A survey of 200 workers on MTurk undertaken by Horton (2011) to investigate their perceived working conditions suggests, however, that "online workers view both offline and online employees more or less equally. In other words, they believe their chances of being treated fairly are as good or better online as they are offline" (MIT Technology Review, 2010). The study then suggests that regulation of the online labour market need to be carefully judged.

Since 2012, Amazon has embarked on an effort to verify all Amazon Payments accounts, including those of MTurk workers in light of criticism of declining working conditions of its international workers, but also due to risks of money laundering. This effort led to the deletion of many MTurk accounts (Ipeirotis, 2013). Furthermore, requesters are now restricted to entities based in the United States (Amazon, 2014a), and only workers in the United States and India can directly access the money transferred to their account, whereas other international workers can only receive the payment in the form of an Amazon gift card (Amazon, 2014b). As a result, MTurk workers, although an internationally diverse group of users, are mostly living in the United States and India today (Ipeirotis, 2010; Techlist, 2014), and the typical turker is not a person that completes tasks in a developing country for a living (Ross et al., 2010).

Database administrator is another traditional data specialist job at the early stage of the data value cycle. These data specialists use specialised software to store, organise and maintain data, such as financial information and customer shipping records. It is also their responsibility to make sure that data are available to users and also to secure databases and data warehouses from unauthorised access.[26] Database administrators hold well-established positions in firms, and the future outlook is positive: the share of jobs related to database administration has been growing since 1999 (see Figure 6.8) and the number of database administrators estimated to grow by 15% from 2012 to 2022 in the United States (BLS, 2014). This is four percentage points faster than the average growth across all occupations shown.

According to population surveys in the United States, the number of sectors employing one or more database administrators per 10 000 employees has increased over the past nine years (OECD, 2013a). In 2012, the five industries with the largest share of database administrators were financial activities (22 database administrators per 10 000 employees); professional and business services (12); wholesale and retail trade (6); manufacturing (6); and information (5, together with public administration and other services). The share of database administrators in these sectors has also increased significantly in recent years, with a remarkable peak in 2011.[27]

However, there is one trend that may reduce the need for database administrators in the near future despite the intensifying use of data and analytics across the economy, and that is the increasing use of online storage and analytics provided by cloud computing. Cloud computing has been described as "a service model for computing services based on a set of computing resources that can be accessed in a flexible, elastic, on-demand way with low management effort" (OECD, 2014c). It thus makes it less and less necessary to deploy internally managed databases, and as a consequence reduces the need for database administrators. As highlighted in OECD, 2010, cost savings through consolidation of ICT infrastructures is one of the expected benefits of cloud computing, and that includes savings of labour costs. Estimates focusing on software-as-a-service suggest that labour cost savings are the second biggest savings potential of cloud computing, after server software costs (see Voce et al., 2009 and MacManu, 2009).

Actuaries, mathematicians, and statisticians are data specialists that have in common that they use quantitative theories and methods to measure and analyse complex phenomena, including assessing uncertainty and dynamic processes to help businesses and clients develop strategies that maximise the business value under these uncertainties (BLS, 2014). These jobs often involve using statistical methods to collect and analyse data and help solve real-world problems in business, engineering, the sciences or other fields (BLS, 2014). Given the rapidly growing volume of data, these specialist jobs have been highlighted by many observers as the most promising job category in the near future. For example, Hal Varian, chief economist at Google, has been quoted saying that "the sexy job in the next 10 years will be statisticians. And I'm not kidding" (Varian cited by Lohr, 2009). This is supported by BLS (2014) estimates suggesting that mathematicians, actuaries and statisticians will be among the fastest growing occupations between 2012 and 2022 (26%). In the United States the overall share of mathematicians, actuaries and statisticians has been decreasing since 2012 (Figure 6.8), but their relative wages continue to increase (Figure 6.10). This suggests that the United States may be currently experiencing an undersupply of these specialists.

The importance of mathematicians, actuaries and statisticians has notably increased in the past year in several industries across OECD economies – particularly in insurance,

advertising and market research. Mathematicians, actuaries and statisticians represented 1.5% of total employment in insurance in 2013, whereas in advertising and market research these occupations accounted for 0.9%, but the share has almost doubled compared to previous year (from 0.46%). Other industries show important increases in the levels of those occupations as well, although their intensities remain low – such as water collection, treatment and supply, where there was a fourfold increase in the level of intensity to 0.2% in 2013 since 2011.

A careful differentiation should be pointed out here, because employment prospects are not as positive for all mathematicians, actuaries and statisticians. At the 4th *OECD Global Forum on the Knowledge Economy* (GFKE, see Annex of Chapter 1 for the highlights), several participants noted that advanced analytic tools are now able to automate many simple tasks that statisticians used to do manually. Increasingly, these tools can, for instance, automatically fit thousands of statistical models to the available data and automatically generate and test different hypotheses (Davenport, 2014). A statistician relying on manual hypothesis testing can typically create only a few models per week. The increasing capacity of data analytics will most likely lead to less basic statistical work, which could lead to a negative employment outlook for less skilled mathematicians, actuaries and statisticians, including their related associate professionals and clerks (i.e. assistants).

Data scientists: The high-end all-round data specialists

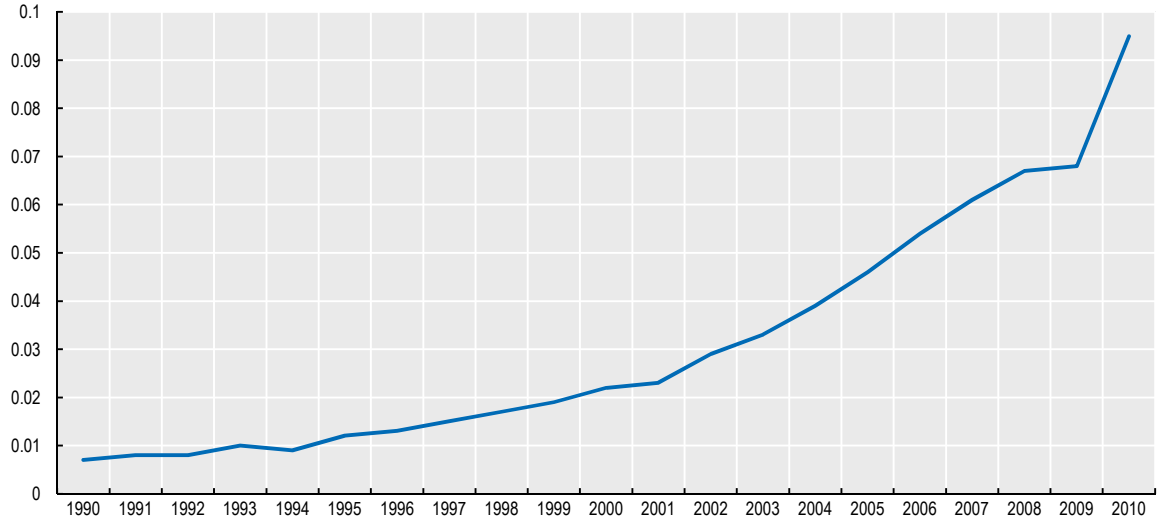
The increase in speed and variety of data-related activities along the data value cycle demands more efficient integration of these working activities, which – thanks to modern ICT tools, including data analytics and cloud computing – has become increasingly cost-effective to undertake. At the same time, business environments increasingly demand that data specialists be flexible and interdisciplinary since they require a great variety of skills, ranging from traditional computer science to mathematics and statistics to domain-specific skills and competence, including those related to business management, finance, marketing and health care, to name but a few. For many data specialists this means that specialist skills limited to one phase of the data value cycle will not be enough in the long run, and that there will be an increasing need to see the bigger picture to tackle all aspects of a problem, from initial data collection and data analysis to drawing conclusions for informing and even automating decision making.

In addition, the convergence of disciplines such as mathematics and statistics – but also, as mentioned above, natural and social sciences such as physics and economics – with computer science has been observed for some time (OECD, 2014d). As convergence has led to the cross-fertilisation of all related disciplines, it has also blurred the boundaries between the disciplines, including biology and computer science in the case of bio-informatics; finance, economics and computer science in the case of computational economics; and statistics and computer science – in particular artificial intelligence – in the case of machine learning, sometimes also referred to as statistical learning (see Chapter 3 of this volume). For (newly trained) data specialists, the convergence of disciplines provides an opportunity to respond to the increasing needs in current business and scientific environments for more flexible and interdisciplinary data-related work. But for many (established) data specialists this also increases the skills requirement considerably, as the use of new and more advanced data analytic tools and techniques is increasingly expected; many of these tools and techniques, such as advanced machine-learning algorithms, are being developed and used in computer science.

The trends presented above (convergence and the cost reduction in data-related activities) have led to the transformation of existing data specialist jobs into a job category or career path that is now commonly referred to as data scientist (Royster, 2013). The job title "data scientist" arose in recent years and has been most frequently used to describe professionals working on big data projects; however, there still is no widely agreed definition. The specific role of a data scientist is diverse and cannot be easily generalised, since a data scientist is expected to be an *all-round data specialist*. This has made measuring the number of data scientists particularly challenging; efforts to do so now rely on surveys such by King and Magoulas (2013, 2014) or analyses of social networks such as those of Patil (2011) and Tambe (2014). Analysis of job starters on LinkedIn by Patil (2011), for instance, shows that the number of persons embarking on careers as data scientists is growing exponentially, although the level remains low at 0.1% of all job starters in 2010 (Figure 6.15). Although they show a clear growing trend, these statistics have one serious limitation: they rely on workers to accurately classify themselves as "data scientists". In addition, Tambe (2014) points out that "there is a potential bias in online platform participation towards younger workers who use emerging technologies. Older IT workers using mature information technologies may have less incentive and lower proclivity to post their technical skills on LinkedIn".

Figure 6.15. **Growth of job starters listed in LinkedIn with a focus on data analytics and data science**

As percentage of all job starters



Source: Patil, 2011, based on LinkedIn.

Looking at various definitions used in the past years to characterise data scientists provides first insights into the occupations as well as the tasks and skills related to data scientists. Originally the term data scientists was coined by statistician Jeff Wu (1998, cited by Kuonen, 2014), who claimed that statisticians should be called data scientists since they spend most of their time manipulating and experimenting with data. In 2005, the US National Science Foundation (NSF) published a report that called for data scientists who are: "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection" (NSF, 2005, Chapter 3). DJ Patil and Jeff Hammerbacher, who

work with big data in LinkedIn and Facebook, respectively claim to have further developed the job title "data scientist" in 2008 in order to describe their position as "high-ranking professional with the training and curiosity to make discoveries in the world of big data" (Davenport and Patil, 2012).

The term, however, is somewhat vague. Many authors highlight the combination of statistics, programming and data visualisation skills as a key differentiating factor. Kenneth Cukier notes in *The Economist* (2010) that data scientists are a new kind of professional who combines the skills of software programmers, statisticians and storytellers/artists to extract insights from data. Similarly, Davenport (2012) calls data scientists "magicians who transform an inchoate mass of bits into a fit subject for analysis." In his opinion data scientists can extract data out of a server log, a telecom billing file, or the alternator on a locomotive, and analyse it. They also create new products and services for customers and also interact with senior executives and product managers (Davenport, 2012). According to the Data Science Association (2012), "data science" means the scientific study of the creation, validation and transformation of data to create meaning, and a "data scientist" is a professional who uses scientific methods to liberate and create meaning from raw data. Furthermore, they often work as data visualisers to create visualisations using both open source and proprietary tools to communicate their findings (UN Global Pulse, 2013).

In some cases, basic programming skills are not enough; what are needed are advanced (software) engineering skills, including expertise in machine learning (ML). Bertolucci (2014), Brave (2012) and the UK National Career Service (2014), for instance, stress that data scientists will often be in charge of integrating data from a variety of sources (i.e. data mashups) in order to develop data-driven applications building on ML algorithms. To do this data, scientists need to create and maintain complex software systems based on big data-specific technologies like Hadoop, Hive, Pig, HBase and Cassandra (Insight Data Engineering, 2014). Most of these technologies are so new however that few experts have sufficient knowledge or the expertise to work with them, and those with high levels of skills tend to concentrate in specific regions. Analysis of LinkedIn profiles by Tambe (2014) suggests that expertise in Hadoop, a major big data-related technology, is concentrated in certain regions in the United States, with the San Francisco Bay area being the most Hadoop-intensive region. His analysis underlines that geography matters for unleashing labour market spillovers, and provides an explanation for the systematic cross-regional firm-level variations in IT returns observed by many authors, such as Brynjolfsson and Hitt (2000), Dewan and Kraemer (2000), and Bloom and Van Reenen (2007). But these findings also call for a cautious interpretation of country-level employment statistics, which do not reflect (sub-)regional labour market concentrations and dynamics.

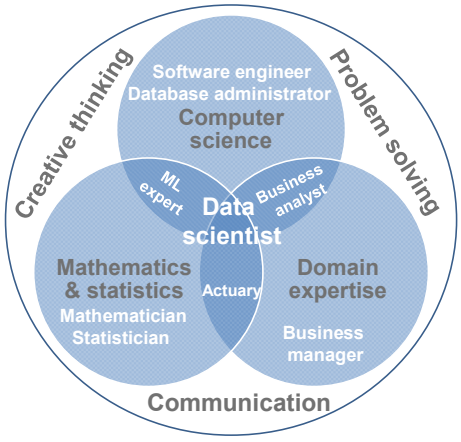
Last but not least, authors highlight that data scientists have domain-specific competence. Many authors stress in particular business-related skills and competence. Brave (2012), for instance, highlights that data scientists need to be able to analyse data sets to extract the domain- or business-relevant information. IBM (2014) stresses business acumen as a key ability of data scientists, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organisation approaches a business challenge. Brynjolfsson and McAfee (2012b) also highlight that data scientists have a high commercial awareness and knowledge of business processes to help decision makers reformulate their challenges so that big data can tackle them.

It should be noted that the need for domain-specific competence is not new, and has been often highlighted in the past with respect to ICT skills (see e.g. OECD, 2005 and 2012b). This also raises the question of the extent to which "data scientists" is a new job category. Some have criticised the current data scientist debate as overhyped, pointing to similar exaggeration around "big data". Harris, Murphy and Vaisman (2013), for instance, believe that the job title data scientists is more a buzzword then a real job description, since "the people doing this work used to come from more traditional and established fields: statistics, machine learning, databases, operations research, business intelligence, social or physical sciences, and more." In line with this, Kuonen (2014) believes that the occupation of a data scientist is not very different from what was formerly attributed to a statistician, since data have long played a role in advising and assisting operational and strategic thinking. For him as statistician, the term "data science", which is often used synonymously with data mining (Provost and Fawcett, 2013), is a part of statistics.

The advocates of data science, in contrast, claim that although statistics are an important part of data science, many of the key techniques for using big data are rarely taught in traditional statistics courses (Brynjolfsson and McAfee, 2012b). Kuonen (2014) also admits that there is a difference between data science and statistics. Traditional statistical analysis focuses on experimental data analysis and the testing of hypothesis, and thus takes a top-down approach; data science focuses on analysing observational data and aims at discovering new ideas with a bottom-up approach. Gartner researchers (Laney, 2012) found that data scientists are expected to work more in teams and to be more skilled at communication compared with traditional statisticians. They also frequently require experience in machine learning, computing and algorithms, and are required to have a PhD nearly twice as often as statisticians. Even the technology requirements for each role differ, with data scientist job descriptions more frequently mentioning Hadoop, Pig, Python and Java among others (Laney, 2012). Loukides (2011) summarises the discussion by explaining that the occupation of a data scientist is closely related to traditional occupations that require a strong mathematical background and computing skills. This is especially true since most data scientists on the job come from a discipline in which "survival" depends on getting the most from the data, such as physics, statistics and economics.

All these definitions and controversial debates presented above underline that data specialists have increasing skill requirements, with data scientists being – if not a "new" job category – at least the most advanced and talented data specialists. A transformation of data specialist jobs towards data scientists can also be observed. In other words, data specialists will increasingly need to combine skills and competences needed to collect, analyse, and use data across the data value cycle in a way that clearly creates value added for their organisation. In particular, data specialists will typically be required to have a mix of different skill sets, including computer science skills such as software engineering, database management, and machine learning (ML), as well as skills in statistics and domain-specific skills such as business management, marketing, finance and health (Figure 6.16). Data specialist skills therefore are not limited to (traditional) ICT specialist skills, although ICT specialist skills such as programming and database administration provide the basis for many future data specialist jobs including data scientists. In addition, "soft skills" such as communication, creative thinking and problem solving skills are also often increasingly highlighted as skill requirements (see discussion in previous section).

Figure 6.16. **Data specialist skills and competence mix**

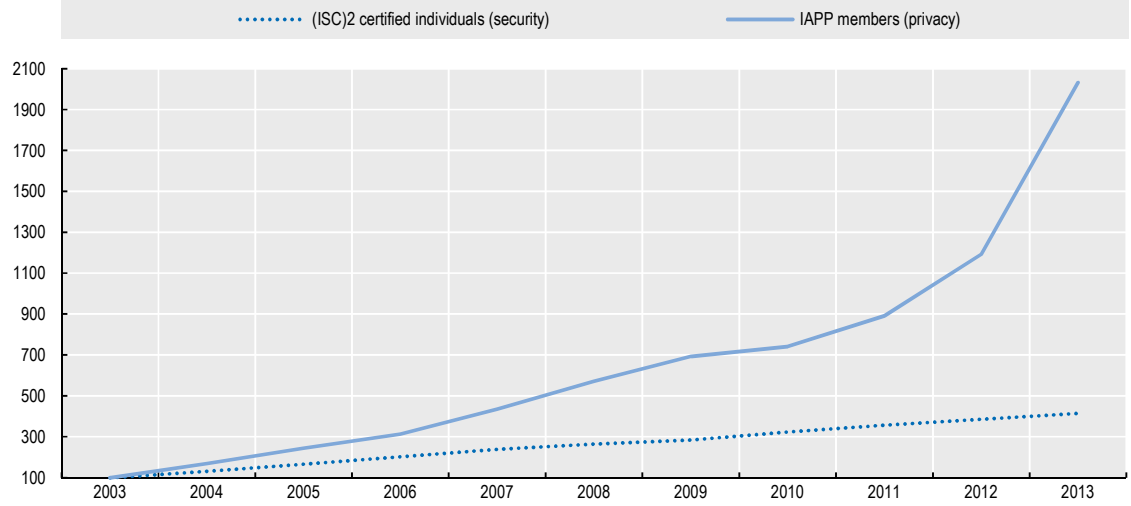


Security and privacy professionals

For data-centred organisations, meeting privacy and security expectations requires more than legal compliance and sound security practices (see Chapter 5 of this volume). Under the 2013 revisions to the OECD Privacy Guidelines, for example, accountable organisations need to put in place multifaceted privacy management programmes, and stand ready to demonstrate them on request from a privacy enforcement authority (OECD, 2013c, paragraph 15). The range of skills needed to implement such programmes is broad, covering legal, technical, communications, governance and public relations aspects, for example. This has increased the need for experts in security and privacy. The steady growth in demand for security expertise seen over the past decade continues, while for privacy professionals the growth has been rapidly accelerating in recent years (Figure 6.17). For both security and privacy, the difficulty in finding available professionals with the required skills and expertise remains a challenge for organisations looking to strengthen their capacities in these areas.

Figure 6.17. **Trends in the number of certified/professional privacy and security experts, 2003-13**

Index, 100 = 2003



Source: Based on annual reports of the International Information Systems Security Certification Consortium [(ISC)^2, 2011] and of the International Association of Privacy Professionals (IAPP, 2013; 2014).

Cybersecurity professionals

The number of cybersecurity professionals worldwide continues to rise steadily, as evidenced by the growing number of individuals with professional certifications for cybersecurity skills, such as the International Information Systems Security Certification Consortium [(ISC)^2] (Figure 6.17). As of the end of 2013, (ISC)^2 had certified 95 781 individuals worldwide, representing a fourfold increase over a decade. In the United States, the Bureau of Labor statistics shows that demand for graduate-level cybersecurity workers will rise by 37% over the next decade – more than twice the predicted rate of increase for the computer industry overall.[28]

Despite this rise, the supply of skilled cybersecurity professionals falls well short of demand. A 2013 report by Japan's National Information Security Center suggested a shortage of 80 000 information security engineers in the country (Humber and Reidy, 2014). Moreover, the report argued that most practicing cybersecurity professionals lack the skills necessary to effectively counter online threats.[29] Likewise, in the United Kingdom an analysis of official statistics on students leaving higher education in 2012-13 showed that less than 1% of computer science graduates in employment were in cybersecurity roles.[30] The National Audit Office of the United Kingdom has warned that it could take another 20 years to tackle the skills gap in trained cybersecurity staff.[31] The government has reacted to address this skills shortage. Specifically, the National Cyber Security Strategy of the United Kingdom aims to develop crosscutting knowledge, skills and capability. Through the National Cyber Security Programme, the Department for Business Innovation and Skills, the Government Communications Headquarters and the Cabinet Office have partnered to lead and support activity to increase cybersecurity skills at all levels of education.[32]

Privacy professionals

One of the more important developments to improve the effectiveness of privacy protection measures has been the emergence of a professional class of privacy experts in organisations (Bamberger and Mulligan, 2011). In some cases there is a statutory basis to support or encourage the role of the privacy professional. For example, Germany's Bundesdatenschutzgesetz (Federal Data Protection Act) sets out specific requirements concerning data protection officials in organisations. Canada's federal private sector legislation, PIPEDA, requires an organisation to designate an individual(s) to be responsible for its personal data-handling activities. New Zealand's Privacy Act requires every agency in both the public and private sectors to appoint a privacy officer, and various pieces of US legislation require federal agencies to have Chief Privacy Officers or Senior Agency Officials for Privacy. Furthermore, the current EU Privacy Directive also contains a reference to a personal data protection official, and the proposed EU data protection regulation would require that data protection officers be appointed for all public authorities and for companies processing more than 5 000 data subjects within 12 months. This would further elevate the numbers of professionals.

The growth in the number of privacy experts has also been encouraged and supported by professional associations, setting the scene for the development of a privacy workforce, including chief privacy officers (CPOs) (Clearwater and Hughes, 2013). These associations provide training, certification, conferences, publications, professional resources and industry research to a growing membership. The largest and most global in reach – the International Association of Privacy Professionals (IAPP) – now has more than 18 000 members (a 24% increase from September 2013) in 83 countries around the

world (Figure 6.17). Other associations include the Privacy Officers Network, through which senior privacy officers involved in the practical implementation of privacy initiatives meet and exchange ideas through a professional support network,[33] and national bodies like the Association Française des Correspondants à la Protection des Données à Caractère Personne (France)[34] and the Asociación Profesional Española de Privacidad (Spain).[35]

The steep growth in IAPP's membership numbers – from 10 000 members in 2012 to a projected 20 000 at the end of 2014 – demonstrates the broad recognition in the marketplace of the importance of sound data governance practices. In its Fortune 1000 Privacy Program Benchmarking Study, the IAPP documents that while budgets vary widely across Fortune 1000 companies, the average privacy budget is USD 2.4 million, 80% of which is spent internally on areas ranging from developing policies, training, certification and communications, to audits and data inventories. The study also highlights that privacy budgets are likely to grow, with nearly 40% of privacy professionals predicting an increase in their budget in the coming year (by an average of 34%) and 33% intending to hire new privacy staff. The IAPP's annual salary survey corroborates the results of the benchmarking study. The survey continues to demonstrate a steady increase in privacy officers' pay, with CPOs earning an average of USD 180 000 per year in the United States, and privacy leaders (who do not hold the title of CPO) earning an average of USD 131 000 in the United States, and USD 125 000 worldwide.[36]

Although the growth in security and privacy professionals documented in this section is both impressive and important, it does not fully capture the move to more deeply devote attention to these topics across workflows in some organisations. For these organisations, such issues are seen not just as the responsibility of designated privacy/security staff, but as a shared responsibility across the parts of the organisation that deal with personal data or matters impacting security. In particular, as companies move beyond viewing privacy as a compliance matter to be addressed by legal departments or as a technical issue handled by IT departments, they will need to put in place ethical review processes and ensure that they have privacy- and security-literate employees.

6.3. Promoting data-driven innovation and smoothing structural change

Seen against the background of one of the most important technological breakthroughs and innovation processes in human history, DDI carries great potential for improving the future of humankind in a world of global challenges, including unsolved development challenges. DDI is a new source of growth that can boost the productivity and competitiveness of all industries and economies. However, it is also disruptive, with a potential for "creative destruction" within labour markets that requires permanent and careful observation by policy makers. That effort is all the more necessary due to the potential of DDI to render many intellectually demanding jobs obsolete, including some data specialist jobs.

To what extent and within what time horizon DDI may lead to "technological unemployment" is still open and subject to academic debates.[37] But there is evidence that it may further increase inequality in earnings through skill-biased technological change, if not addressed by measures implemented through (inter alia) social and tax policies. In the current context of weak global recovery and lingering high unemployment in major

advanced economies, the associated risks of unemployment and inequality in earnings thus deserves policy makers' attention.

The discussion in this section stems from the premise that "Resilient Economies and Inclusive Societies" is the yardstick to orientate policies for promoting jobs and "Environmentally Sustainable ('Greener') Growth", as underlined by Ministers and Representatives[38] in the OECD (2014a) Ministerial Council Statement. The following sections suggest that a "double strategy" is needed that i) supports the development and strengthening of the right mix of skills and competencies needed, including but not limited to data specialist skills, and ii) promotes social cohesion while addressing the risk of inequality that is rising in a number of countries (Herlyn et al., 2015). In the context of DDI, inequality could become a major issue, especially if access to high-quality education, which is urgently needed (before and after a first entry into the labour market) to take advantage of the job creation opportunities ahead, is limited to few (see Cingano, 2014).

Satisfying skill and competency needs

Since the first Industrial Revolution, the education system has been key in supporting the need for structural adjustment in a "race between education and technology" (Goldin and Katz, 2008). For a long time and up until now, most societies have been successful in that race: new and better jobs for a better educated workforce have been the result, although structural adjustment often took time to take place and was not always very inclusive.

In light of the major implications of DDI discussed in this chapter, humans will need a broad education as a basis for the accelerating race ahead. An overly narrow education system, trimmed down to the specific jobs requirements of a particular time, will most likely not be the robust strategy needed to meet the skill challenges described in this chapter. The discussion presented here rather suggests that humans need to further develop their competitive advantage over machines and refine the skills that machines will not be able to perform at the upper end. This includes the development of a broader interdisciplinary understanding of multiple complex subjects, but also deeper insights into some domain-specific issues. A solid intellectual foundation in STEM (science, technology, engineering and mathematics), including in particular statistics and computer science, is necessary but not sufficient. At least as important is a true understanding of social and legal systems – in particular, of economics, ecology, human behaviour, legal requirements (e.g. relating to privacy and intellectual property rights), and – last but not least – ethics.

Creative thinking, problem solving, and communication skills have to be strengthened and cultivated as well as sensomotoric skills, as these will be the skills through which humans will outperform machines for a long time. In that respect humans are the best "combination" of abstract abilities and impressive sensomotoric skills, and this remains the key opportunity for job creation in the long run. But this also means that our "body and mind" need to be trained and kept in good shape throughout our lives. Achieving these huge education requirements will not be easy for individuals or for national education systems. Current pressure towards shorter time for education may make it more challenging to develop the required skills and competencies needed for a more inclusive "race between education and technology". In some cases formal education institutions may not be best placed to provide all the necessary skills. Therefore, efforts towards lifelong learning need to be supported by all stakeholders of the education system,

including parents, formal education institutions, businesses, labour unions and governments. This requires a strategic approach for strengthening skill systems, as highlighted by the OECD Skills Strategy (OECD, 2012a) (see Box 6.7).

This section discusses means for developing the relevant data specialist skills, with a particular focus on formal education institutions. Further work would be required to fully apply the OECD Skills Strategy framework to assess how to better activate the supply of the skills needed for the data driven economy and how to put these skills to more effective use. Moreover, further reflection is needed to better understand how to better develop the competitive advantages humans have over machines.

Box 6.7. **The OECD Skills Strategy**

The OECD (2012) Skills Strategy framework provides countries with a strategic approach to strengthen their skills system in building, maintaining and using their human capital to boost employment and economic growth, and promote social inclusion and participation. It encompasses the following objectives:

1. Developing relevant skills, by i) encouraging and enabling people to acquire the right skills throughout life, ii) fostering international mobility of skilled people to fill skill gaps, and iii) promoting cross-border skills policies
2. Activating skills supply, by i) encouraging people to offer their skills to the labour market and ii) retaining skilled people in the labour market
3. Putting skills to effective use, by i) creating a better match between people's skills and the requirements of their job and ii) increasing the demand for skills.

Source: OECD, 2012a.

Developing the basic skills needed

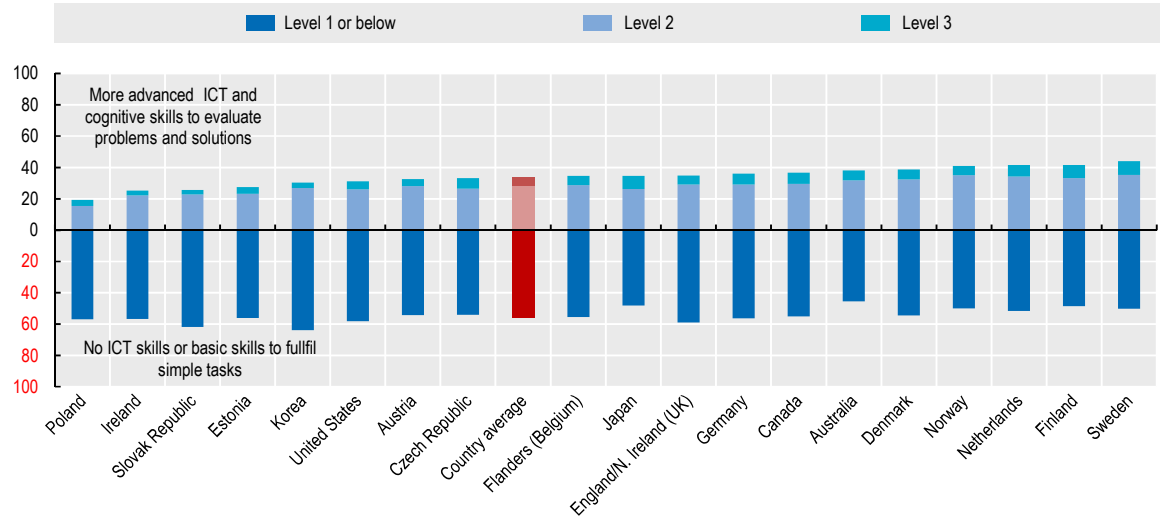
In the past there have been considerable mismatches between the supply of and demand for ICT skills in general and software skills in particular (OECD, 2012a). Shortfalls in domestic supply, owing to a large share of students leaving compulsory education; lack of educational courses and little vocational education or on-the-job training in industry; and the low share of female employees in ICT specialist occupations were often highlighted as factors limiting the availability of ICT specialist skills. And this could remain true for data specialist skills. Furthermore, restrictions on the immigration of highly skilled personnel and difficulties in international sourcing of analytical tasks – which could intensify due to current considerations of data localisation requirements – put further pressure on national education systems to develop the right mix of skills needed for the data-driven economy.

Besides language and reasoning skills, creativity and social intelligence, and perception and sensomotoric skills, basic ICT literacy needs to be much further developed, given that it has become the skill foundation of the workforce in the digital economy; this includes knowledge about security and privacy risks, as specifically identified in the revised OECD Privacy Guidelines' call for "complementary measures". However, an OECD (2014g) study based on data from the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) reveals that 7% to as much as 27% of adults have no experience in using computers, or lack the most elementary computer skills, such as the ability to use a mouse. The study also shows that in OECD

countries, only 6% of the population is categorised with the "highest level" of ICT skills, meaning "they can complete tasks involving multiple applications, a large number of steps, impasses, and the discovery and use of *ad hoc* commands in a novel environment". In countries such as Austria, the United States, Korea, Estonia, the Slovak Republic, Ireland and Poland, the share is 5% and below (Figure 6.18). This suggests that for most OECD countries, the basis for developing data specialist skills is very weak.

Figure 6.18. **Level of proficiency in problem solving in technology-rich environments, 2012**

As a percentage of 16-65 year-olds



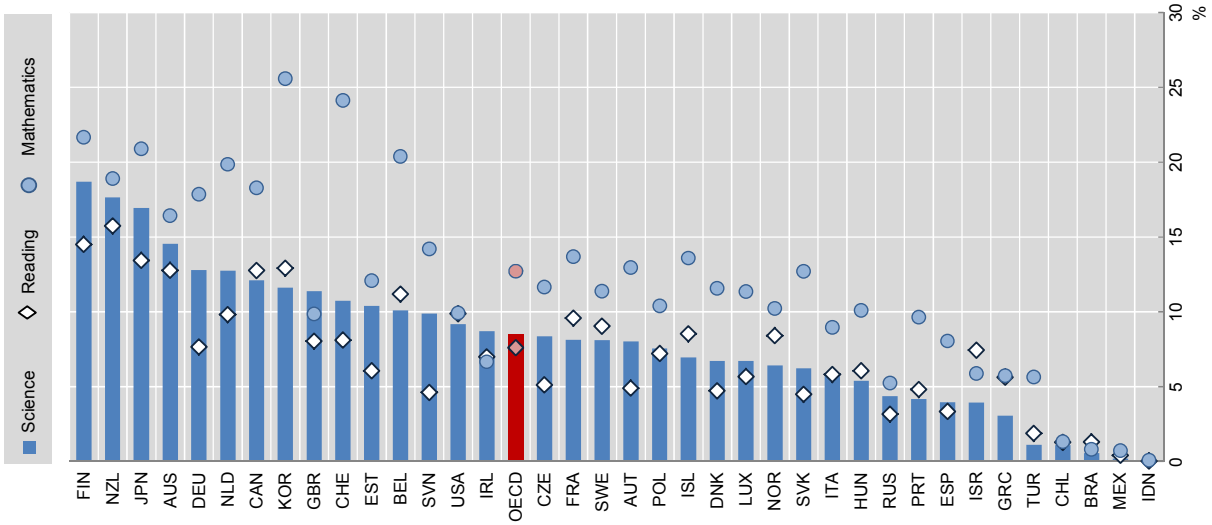
Note: Problem solving in technology-rich environments requires "computer literacy" skills (i.e. the capacity to use ICT tools and applications) and the cognitive skills required to solve problems. Level 1 or below possesses no ICT or basic skills to fulfil simple tasks; levels 2 and 3 require more advanced ICT and cognitive skills to evaluate and find solutions.

Source: *OECD Science, Technology and Industry Outlook 2014*, based on OECD's Programme for the International Assessment of Adult Competencies (PIAAC), http://dx.doi.org/10.1787/888933151932.

The discussion presented in this chapter highlighted mathematics proficiency as an important foundation for data specialist skills besides language and reasoning skills. Mathematics and statistics prepare students to work with data analytics. "Math helps students develop the logical thinking and problem-solving skills they need. Statistics provides the analytical knowledge that they need to properly study the data and to interpret the results in a meaningful way" (Royster, 2013). That means that countries with a high share of top-performing students in mathematics but also reading and science are more likely to develop talent pools for future data specialists. Results from the 2009 OECD Programme for International Student Assessment (PISA) on the science, reading and mathematics proficiency of 15-year-olds show that 13% of students in the OECD area were top performers in mathematics, 9% in science, and 8% in reading (Figure 6.19). The share of top performers in mathematics is highest in Korea, Switzerland, Finland, Japan and Belgium. It is the lowest in Mexico, Chile, Turkey, Greece and Israel, besides partner economies Indonesia, Brazil and the Russian Federation. These low performers could be facing difficulties in developing data specialist skills the next five to ten years.[39]

Figure 6.19. **Science, reading and mathematics proficiency at age 15, 2009**

Percentage of top performers



Source: OECD Science, Technology and Industry Scoreboard 2013, http://dx.doi.org/10.1787/888932890675.

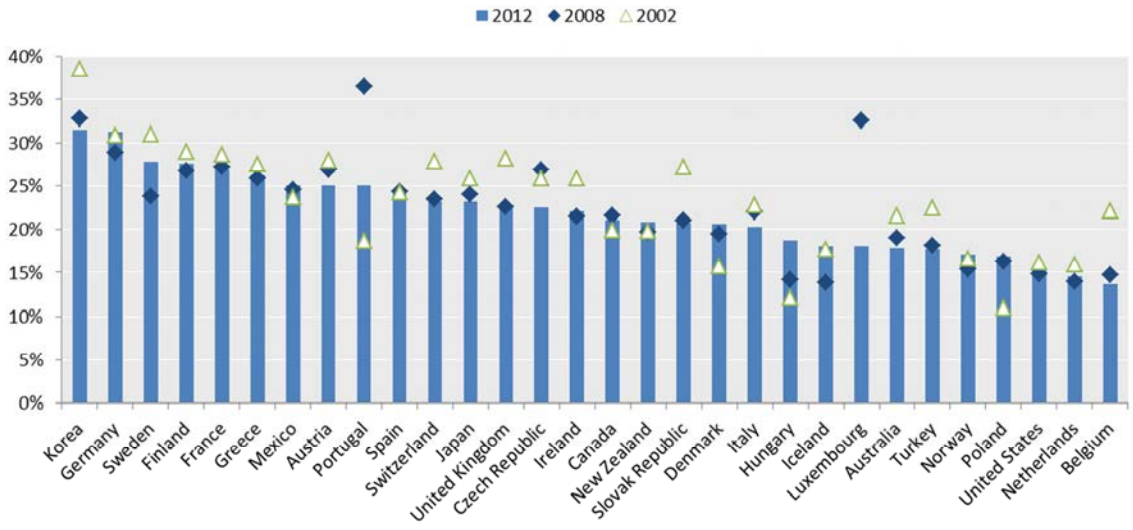
Developing the higher skills needed

Higher education institutions have a pivotal role to play in providing the higher-level skills needed by data specialists. This is especially the case for computer science skills, since very few secondary education institutions provide skills such as programming and database management. These skills are essential however, since they not only provide the ability to develop, maintain, and operate data-driven systems, but they also train logical thinking and problem solving (Royster, 2013). Evidence suggests that the supply of computer science graduates is progressing. Over a period of five years (2000-05), the number of computer science graduates in OECD countries for which data are available almost doubled, but then started to decline until 2010. Since then, the number of graduates in computer science has been increasing, reaching almost the same level in 2012 as in 2006.

But in addition to a degree in computer science (often with an emphasis on data management, data mining or artificial intelligence), a significant number of data specialists have a degree in experimental physics, molecular biology, or bio-informatics, which often involves analysis of large data sets (Loukides, 2011; Rogers, 2012). This suggests that data specialist skills are not only provided in computer science study programmes, but they are also more likely to become an integral part of science, technology, engineering and mathematics (STEM) disciplines, in particular given the convergence of the disciplines with computer science as highlighted above and the increasing data intensity of science and engineering (see Chapter 7 of this volume). However, the share of all STEM graduates from OECD countries for which data are available declined from 22% in 2000 to 21% in 2012 (see Figure 6.20), and the participation of women in tertiary-level graduate STEM programmes remains low (see Box 6.5). These trends indicate a long-term stagnation of the relative supply from high-demand science and technology oriented fields. In many countries in 2012, the supply of higher education graduates from STEM-related study fields indeed stagnated (Norway,

Ireland, Poland and Greece) or even declined (Czech Republic, Korea, France, Slovak Republic, Spain and Italy) in absolute numbers. But the decline is also due to a faster growth in the number of graduates in non-STEM fields (see e.g. Austria).

Figure 6.20. **STEM (science, technology, engineering and mathematics) graduates**
As percentage of total graduates



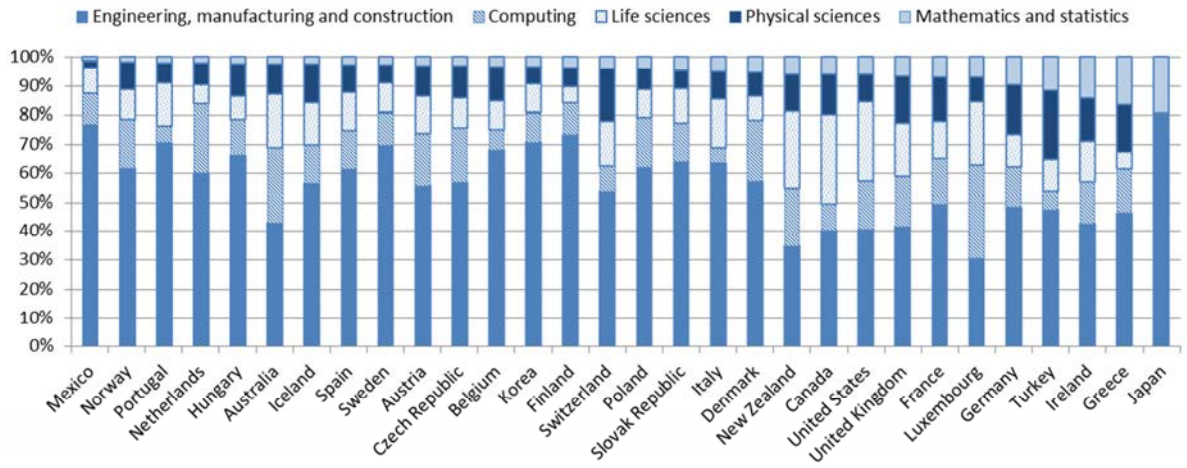
Note: Graduates are those who successfully complete an educational programme during the reference year of the data collection. The year shown is the year in which students graduate, with the exceptions of Denmark, Finland, France (until 2002) and Italy where students graduate the previous year.

Source: Based on UNESCO-OECD-Eurostat (UOE) data collection on education statistics, compiled on the basis of national administrative sources, reported by Ministries of Education or National Statistical Offices.

A closer look at STEM graduates by disciplines is worth taking, in particular the share of mathematics and statistics graduates. Not only are there huge cross-country variations, but when considering that mathematics and statistics are most likely the most demanding data specialist occupations, with indications of shortages in particular in the United States, it is striking to see these disciplines account by far for the lowest share of STEM graduates in some countries (between 2% in countries such as Mexico, Norway, Portugal, and the Netherlands, and 10% in Germany, 12% in Turkey, 14% in Ireland, and 17% in Greece) (Figure 6.21).

At the same time, STEM should not be overrated, since they are not the only source for the skills and competences needed in a data-driven economy. As highlighted above, a solid intellectual foundation in STEM is not sufficient. But even when focusing on data specialist skills, one can observe that these skills may even be acquired in disciplines beyond STEM. For example, the emergence of trends such as "data journalism", where journalism is centred on the use data and data visualisation, suggests that data analytic skills may also be part of tomorrow's curricula for a wider range of study programmes including journalism.

Figure 6.21. **STEM graduates by disciplines, 2012**



Source: Based on UNESCO-OECD-Eurostat (UOE) data collection on education statistics, compiled on the basis of national administrative sources, reported by Ministries of Education or National Statistical Offices.

In order to meet the increasing demand for data specialists, an increasing number of education institutions are offering specialised master's-level programmes related to data and analytics, some of which are advertised as "data science" programmes (Violino, 2014). In the United States, a growing number of data science programmes, often master's-level programmes, can be observed. Northwestern and North Carolina State University offer, for instance, a Master of Science in Analytics (MSiA) programme that provides students with "coursework in statistics, modeling, operations research, quantitative analysis, decision analysis, databases, and data management form the core of the curriculum".[40] Most programmes offer a similar curriculum. Few however provide domain-specific skills and competence such as business and entrepreneurial, and communication skills in addition. And even less provide coursework on privacy protection and cybersecurity.

The very new and rapidly changing nature of ICTs including data analytics makes workplace training, in addition to formal education, increasingly important for augmenting and adapting workers' skills. As Royster (2013) confirms, "even workers who have a statistics or data analysis background need to stay current with the fast-changing world of big data". This is especially true for older workers, for whom skills acquired through the educational system are likely to be missing or substantially depreciated in the field of data specialist skills. Evidence in the past suggested that the share of firms with vocational training on ICTs decreased between 1999 and 2005, which could raise concern in policy makers. Although no representative statistics exist, there is anecdotal evidence in the case of data analytics that in particular large firms are offering data analytic training for workers who have little formal training. Many data scientists training programmes developed by private sector companies such as IBM are also offered to the public (Marsan, 2012). In addition, Massive Open Online Courses (MOOCs) on data analytics are increasingly being offered. For instance, in February 2015 Coursera, one of the leading MOOC providers, offered 69 courses related to "Statistics and Data Analysis". This was by far the most frequent course category in Coursera, before "Information, Tech & Design" (25), "Social Sciences" (15) and "Mathematics" (14).

Addressing the risks of rising income inequality

Beside policies aiming at enhancing skills, further policy consideration may be needed to smooth structural change and in particular address the inequality that could increase due to DDI. A recent report by Cingano (2014) provides evidence suggesting that "reducing income inequality would boost economic growth" (OECD, 2014e). The study suggests that economies where income inequality is decreasing grow faster than those with rising inequality. The study underlines that lack of investment in education is the main mechanism through which growth is hampered due to inequality: in particular for children from poor socio-economic backgrounds, lack of investment in education lowers social mobility and hampers their skills development, thereby negatively affecting growth potentials in the long run. The study therefore highlights the need "to promote equality of opportunity in access to and quality of education" (Cingano, 2014). But the study also stresses that "redistribution policies via taxes and transfers are a key tool to ensure the benefits of growth are more broadly distributed and the results suggest they need not be expected to undermine growth" (Cingano, 2014).

As highlighted above, DDI offers potentially significant increases in labour productivity through the automation of cognitive and manual tasks, many of which are high value added tasks considered to be not susceptible to machines automation (i.e. taxi drivers affected by driverless cars; research clerks and assistants affected by expert systems such as IBM Watson). Therefore, several authors including Ford (2009), Cowen (2013), Cukier and Mayer-Schönberger (2013), Frey and Osborne (2013), Levy and Murnane (2013), Brynjolfsson and McAfee (2014), Rifkin (2014) and Elliott (2014), have highlighted the potential negative implications of data-driven automation on wage and income inequalities.[41] Brynjolfsson and McAfee (2012a; 2014), for instance, refer to work by Acemoglu and Autor (2010) on "skill-biased technical change" to present evidence on the growing relative wages of higher skilled workers compared to lower skilled workers. This trend could intensify with DDI, which tends to amplify employment polarisation as middle income jobs become more susceptible to machine automation as highlighted above.

Some of these implications were highlighted already in the OECD (2013b) Report on *Supporting Investment in Knowledge Capital, Growth and Innovation*, which looked at the policy implications of knowledge-based capital (KBC), and concluded that:

> *KBC-based economies rewards skills and those who perform non-routine manual and cognitive tasks, but may also reward investors (who ultimately own much of the KBC) over workers (in the United States, for instance, wages as a share of GDP are at an all-time low). Rising investment in KBC can create winner-takes-all opportunities for a few, while entire occupational categories can be replaced by machines and software. (OECD, 2013b)*

OECD (2013b) points to a much broader set of issues about the remuneration of labour versus remuneration of capital,[42] that many of the authors listed above have also called attention to in light of the still open academic debate about "technological unemployment". Their main argument can be summarised as follows: if the need to sufficiently pay a highly educated workforce should be eliminated via smart applications that can (partly) perform knowledge- and labour-intensive tasks with less labour, including less educated and paid labour, then the known pattern of income distribution could be in danger. Often, redistribution policies including also (i) negative income tax (NIT) and (ii) an unconditional basic income, are then proposed for consideration. NIT, such as proposed for example by Friedman (1962), provides citizens whose income is

below a certain threshold with supplemental payments from the government (instead of paying taxes to the government). Some have suggested that earned income tax credits (EITC) already provided for example in the United States to low or moderate income working individuals would be a policy measure similar to NIT (Farrell, 2013). Unconditional basic income, in contrast, is paid to every citizen either employed or unemployed. It is sometimes suggested that its amount should reflect national threshold definitions under which citizens would be categorised as "poor". The impact of both policy instruments and the challenges they raise are still subject to controversial debates, as is the question about "technological unemployment". That said, inequality will remain high on policy makers' agenda, in particular if the availability of natural resources remains a huge constraining factor.

6.4. Key findings and policy conclusions

DDI can be disruptive with the potential to amplify employment polarisation as it affects a broader range of middle income jobs – the segment of the population that glues our societies together. Moreover, DDI can contribute to employment polarisation through (i) data-driven decision automation affecting white collar jobs, and (ii) the enabled new generation of autonomous machines affecting blue collar jobs in particular in manufacturing and logistics. DDI may even negatively affect some data specialist jobs such as data entry clerks, database administrators, and statisticians.

That said, it is important to stress that DDI should not be seen in isolation from the rest of the economy. In particular, the dynamics induced by DDI need to be carefully studied. As technological change tends to increase productivity, it reduces costs and increases demand for goods and services, which in turn can help generate more jobs in other parts of the economy. And this is also true for the new goods and services that DDI directly enables. Further studies taking into account these dynamics are therefore needed in order to better understand the employment effects of DDI.

What this chapter clearly showed however is that further investments in education are needed in order to promote the adoption of DDI across society, and to support the developments of the right mix of skills and competence needed for the (re-)creation of decent jobs. The chapter pointed to current debates about the risks that DDI could worsen wage and income inequalities as it amplifies employment polarisation. Evidence shows that economies with decreasing income inequality have grown faster than those with rising inequality (Cingano, 2014). Supporting the development and strengthening of the right mix of skills and competencies, while promoting social cohesion and addressing the rising inequality, could be part of a "double strategy" to prepare societies for the future (Herlyn et al., 2015).

The chapter presented evidence on the data intensification of OECD economies. Data specialist skills are increasingly in demand across industries, although the share in total employment tends to remain low. However, there are signs of an insufficient supply of (basic) skills related to ICT and STEM more broadly, with indications of actual skills shortages with respect to statisticians, mathematicians and actuaries. Developing mathematic and statistic proficiency is essential to appropriately use data and analytics, and how to deal with their limitations discussed in Chapter 3 of this volume.

But the implications of DDI for education are far-reaching, going beyond skills related to ICTs and even STEM. The discussion presented in this chapter suggested that education systems should support a broader interdisciplinary understanding of multiple

complex subjects but also deeper insights into some domain-specific issues. This calls for the development of a solid intellectual foundation in STEM *in combination* with a sufficient understanding of human behaviour and social systems such as provided in humanities. This combination would help enhance qualitative reasoning in addition to quantitative reasoning to enhance the sense of responsibility of future data-informed decision makers.

Furthermore the chapter highlighted that soft skills such as (i) creativity, (ii) problem solving and (iii) communication skills are key for ensuring employment in a data-driven economy in the long run. Furthermore, (iv) highly developed sensomotoric skills will also become a key competitive advantage of humans over machines. These skills, if cultivated with the support of education systems and accompanied by political attention and good co-operative global governance, may lessen concerns related to technological unemployment. This will be more the case if individuals can enhance and complement their talents to use technology to "dance" with the machines instead of "racing" against them.

Annex – Selected statistical definitions of data specialist occupations

Table 6.A1 **Europe: Occupations included in the operational definition of the Data specialists.**

Based on ISCO-08 (3-digits)

ISCO-08	EN Title
212	Mathematicians, actuaries and statisticians
252	Database and network professionals

Table 6.A2 **United States: Occupations included in the operational definition of data specialist**

Based on SOC 2010 (6-digits)

SOC 2010	EN Title	Corresponds to ISCO-08 codes (4-digits)	
15-2011	Actuaries	2120	Mathematicians, actuaries and statisticians
15-2021	Mathematicians	2120	Mathematicians, actuaries and statisticians
15-2031	Operations Research Analysts	2120	Mathematicians, actuaries and statisticians
15-2041	Statisticians	2120	Mathematicians, actuaries and statisticians
19-3022	Survey Researchers	2120	Mathematicians, actuaries and statisticians
15-1141	Database Administrators	2521	Database designers and administrators
15-1142	Network and Computer Systems Administrators	2522	Systems administrators and
15-1143	Computer Network Architects	3514	Web technicians
15-1122	Information Security Analysts	2523	Computer network professionals
15-1199	Computer Occupations, All Other	2529	Database and network professionals not elsewhere classified
		2529	Database and network professionals not elsewhere classified
		2519	Software and applications developers and analysts not elsewhere classified

Table 6.A3 **Australia: Occupations included in the operational definition of data specialist**

Based on ANZSCO 2010 (6-digits)

ANZSCO version 1.2		Corresponds to ISCO-08 codes (4-digits)	
224111	Actuary	2120	Mathematicians, actuaries and statisticians
224112	Mathematician		
224113	Statistician		
262111	Database Administrator	2521	Database designers and administrators
262113	Systems Administrator	2522	Systems administrators
263112	Network Administrator		
263111	Computer Network and Systems Engineer	2523	Computer network professionals
263113	Network Analyst		
262112	ICT Security Specialist	2529	Database and network professionals not elsewhere classified

Table 6. A4 **Canada: Occupations included in the operational definition of data specialist**

Based on NOC 2011 (4-digits)

NOC 2011		Corresponds to ISCO-08 codes (4-digits)	
2161	Mathematicians, Statisticians and Actuaries	2120	Mathematicians, actuaries, statisticians
2172	Database analysts and data administrators	2521	Database designers and administrators

Notes

1 For instance, it is unclear whether those firms adopting DDI became more productive due to DDI-related investments, or whether they were more productive in the first place. Furthermore, these studies rarely control for the possibility that some firms may have eventually seen a reduction in their productivity due to DDI, and as a result may have discontinued their investments in it.

2 For consumer goods and retail firms it is the single biggest barrier, cited by two-thirds of respondents from those sectors. Furthermore, MGI (2011) estimates that the demand for deep analytical positions in the United States could exceed supply by 140 000 to 190 000 positions by 2018.

3 This chapter benefited from, and partly builds on, reflections on the OECD project on data-driven innovation by Estelle L.A. Herlyn, Thomas Kämpke, Franz Josef Radermacher, and Dirk Solte (Research Institute for Applied Knowledge Processing, Ulm, Germany, FAW/n; see Herlyn et al., 2015).

4 With information and communication technologies (ICTs), we have seen the highest innovation speed and greatest penetration rate of new technologies, ever. At the heart of this development is the extreme speed in cost reduction as described by Moore's Law, which holds that processing power doubles about every 18 months relative to the cost or size of central processing units (CPUs) (Moore, 1965). In other words, for decades mankind has been enhancing the performance of processors by a factor of 10 000 every 20 years, which means an improvement factor of more than a trillion over the past 60 years, since the time work on the first transistors or chips started. These are huge achievements – one could easily argue that there has never been so much change induced by technology in such a short time.

The main reason for this explosion of improvement is the possibility of miniaturizing or compressing information. That means that encoding of one unit of information (one bit) requires ever less physical space. This is because the coupling of information and its physical manifestation is very loose. We can make the encoded information (e.g. numbers) always smaller, without changing the results of subsequent algorithmic computations on the information, be it e.g. arithmetic or Boolean operations. That means that in order to add numbers, the size of the physical representation of the numbers is not of principal importance, in contrast to e.g. to physical good such as a car, where the size of the car is essentially a given and not a variable.

5 One way of measuring the IoT is by looking at the number of SIM-cards and phone numbers allocated to M2M communication devices on mobile networks (OECD, 2015). The data show in many countries the market is growing briskly. Most countries report double-digit growth between 2012 and 2013, though most countries do not have data for 2011, so it is hard to analyse trends. Some operators are also reporting on the number of connected devices. AT&T in the United States, for example, reports that it connected 1.3 million devices in the second quarter of 2014, of which 500 000 were vehicles.

6 To make this work, machine learning uses many techniques, also used in data analytics, such as statistical and regression analysis; (unsupervised) learning algorithms and cluster analysis.

7 See the trends referred to as "smart manufacturing", the Industrial Internet, or Industry 4.0.

8 Similar systems underlying ATS are being provided by start-up companies offering online financial advice services at much lower fees than traditional financial advisors (Coombes, 2013).

9 See www.youtube.com/watch?v=sOLXOsiX0Qk on "current and future applications of machine-learning within law, and … automation in the context of legal tasks currently performed by attorneys, including predicting the outcomes of legal cases, finding hidden relationships in legal documents and data, electronic discovery, and the automated organization of documents", accessed 18 May 2015.

10 For example, workers in Amazon's warehouses in the United Kingdom are reported to walk between 11 and 24 kilometres per day (O'Connor, 2013).

11 Before the system can function, it has to model the position of all goods in the warehouse and the most efficient paths and distribution.

12 Unlike Amazon, it cannot move the shelves because the products are bulky and heavy. Instead, it has created a three dimensional storage facility.

13 *Source:* Video clip of Symbotic presentation at www.symbotic.com/robots-in-the-warehouse-expanding-beyond-manufacturing/, accessed 18 May 2015.

14 In this scenario packages are not on pallets, but stacked individually into a delivery van. This again was a computationally hard problem in the past, because it required the packages to be loaded in reverse order from being delivered, but in such a way that the truck can hold the most packages securely. Furthermore, it required three dimensional perception to correctly identify and handle the packages, which are of different sizes and weights.

15 The Amazon warehouse example shows that workers are still necessary to fill orders, to actually pick the products, and to fill the boxes.

16 One sector where autonomous machines are increasingly a reality is the agriculture sector. There are a great number of examples. Algorithms and robots already sort plants such as orchids into various classes and groups, based on optical recognition. Robots harvest lettuce and recognise rotten apples. The spraying of fields is done by tractors that steer themselves and only need minimal operator intervention. Combine harvesters can operate semi-autonomously or work together with a lead harvester. Algorithms vary the spraying of pesticide and fertiliser based on yield data from previous years. The robotic tractor cannot be far off.

17 Together with the MIT economist David Autor, they have examined the changes in occupational distribution in the United States by categorising the work in five areas: 1) solving unstructured problems, 2) working with new information, 3) routing cognitive tasks, 4) routine manual tasks and 5) non-routine manual tasks. The result shows a clear trend, as described in Figure 6.3.

18 It is interesting to note that, given the impressive manifestations of machine intelligence, humans may need to further "re-discover" the importance of their bodies.

19 But he adds that, "the future will bring us The Unaccountable Freestyle Team, The Scary Freestyle Team, and The Crippled Freestyle Team, all at once" (Cowen, 2013, p. 131).

20 Estimates are based on the voluntary, ad hoc module in the EU Community Innovation Survey 2010 on the skills available in enterprises and on methods to stimulate new ideas and creativity. The indicator corresponds to the percentage of firms in the relevant innovation category responding affirmatively to the question: "During the three years 2008 to 2010, did your enterprise employ individuals in-house with the following skills, or obtain these skills from external sources?" Innovative enterprises had innovation activities during 2008-10, relating to the introduction of new products, processes, and organisational or marketing methods. This includes enterprises with ongoing and abandoned activities for product and process innovation. The question on innovation-relevant skills also applies to non-innovative enterprises. Estimates are based on firms with "core" NACE Rev. 2 economic activities (B, C, D, E, G46, H, J58, J61, J62, J63, K and M71).

21 In these countries, the share of data specialists has decreased considerably in recent years. It is important to note that in all countries aside from Greece and Turkey, data specialists mainly comprise database and network professionals.

22 In 2012, information and communication industries accounted for 3.6% of total employment in the OECD area. In nearly all countries, IT and other information services are the largest component of the information and communication industries, accounting for 40%, i.e. 1.4% of total employment in the OECD area. With above 2% of total employment, Ireland was the country with the largest share (2.7%), followed by Luxembourg, the United Kingdom, Finland, Sweden and Denmark. Australia, Greece and Mexico had among the lowest shares, all equal or below 0.5%. Since 2004, the share of IT and other information services in total employment has grown continuously, going from 1.2% to 1.4%. IT and other information services over the last decade have been resilient and driving employment growth, especially during the recent financial crisis where the employment losses were less important than in other industries, with a growth rate in 2009 of less than -1% while the information industries and total employment growth rates were respectively of -3.5% and -2.5%. However, since 2010 IT and information services have been growing quite fast which suggests that ICTs and in particular, IT and information services are playing a significant role in the upcoming recovery.

23 According to data published by the World Information Technology and Services Alliance (WITSA), telecommunications (11.5%), financial services (6.6%), transport (5.1%), health care (4.1%) and government (3.8%) are the five most ICT-intensive sectors. Using ICT intensity as a proxy for data intensity assumes that data-intensive industries have higher ICT expenditure than industries with low data intensity. That assumption can be easily challenged, since data analytics requires less investment in ICTs today (because of cloud computing). In an historical perspective however, this approach can still be useful.

24 Typical examples include text-heavy data sets such as text documents and e-mails, as well as multimedia content such as videos, images and audio streams. Unstructured data account for the largest share of the global data volume by far. According to some estimates, not even 5% of the digital universe can be considered structured or semi-structured data.

25 For example, optical character recognition (OCR) can transform images of text into machine-encoded text, which then can be further processed and used for data analytics, in particular natural language processing (NLP), for tagging or for extracting relevant patterns.

26 Database administrators sometimes share these tasks with network and computer systems administrators, computer network architects, and information security analysts.

27 In 2011, financial activities, professional and business services, information, and public administration were the sectors mainly contributing to the increase in share of database administrators in the United States.

28 See www.bbc.com/news/business-26647795

29 See www.businessweek.com/articles/2014-07-24/proposed-law-would-fix-japans-lax-cybersecurity

30 See www.ft.com/intl/cms/s/0/76b1eef4-1d3c-11e4-8b03-00144feabdc0.html

31 See www.bbc.com/news/business-26647795

32 See https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/289806/bis-14-647-cyber-security-skills-business-perspectives-and-governments-next-steps.pdf

33 See www.privacylaws.com/Privacy-Officers-Network/

34 See www.afcdp.net/

35 See www.apep.es/

36 See 2013 IAPP Privacy Professionals Role, Function and Salary Survey, https://privacyassociation.org/resources/article/2013-iapp-privacy-professionals-role-function-and-salary-survey

37 See Ford (2009), Wilkinson and Pickett (2010), Randers (2012), Cowen (2013), Cukier and Mayer-Schönberger (2013), Frey and Osborne (2013), Levy and Murnane (2013), Brynjolfsson and McAfee (2014), Herlyn and Radermacher (2014), Piketty (2014), and Rifkin (2014).

38 These include ministers from and representatives of Australia, Austria, Belgium, Canada, Chile, Colombia, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Latvia, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom, the United States and the European Union.

39 It is worth noting at this point that new ways to teach math are being developed; see for example Wing, 2008.

40 www.analytics.northwestern.edu/curriculum-and-career-prospects/Course_description.html, accessed 18 May 2015.

41 This leads directly to the question of the implications on DDI on wage and income inequalities, and to the question of the right kind of balance in income concerning the so-called "efficient inequality range" (Cornia and Court, 2001).

42 Cowen (2013, pp. 38-40), for instance, observes that: "In 1990, 63 percent of American national income took the form of payments for labour, but by the middle of 2011 it had fallen to 58 percent." And this trend is not limited to the United States; many OECD countries including France, Germany, and Japan have seen similar trends as highlighted by Piketty (2014). Other authors such as Atkinson (1975), Wilkinson and Pickett (2010), Herlyn (2012), and Herlyn and Radermacher (2014) have discussed the implications more broadly, highlighting for example a trend towards "precarisation" or "neo-feudalisation".

References

Abedi, V. et al. (2012), "An automated framework for hypotheses generation using literature", *BioData mining*, Vol. 5, No. 1.

Acemoglu, D. and D. Autor (2010), "Skills, tasks and technologies: Implications for employment and earnings", *NBER Working Paper* No. 16082, National Bureau of Economic Research.

Amazon (2014a), Amazon Mechanical Turk FAQs, http://aws.amazon.com/mturk/faqs/#Can_international_Requesters_use_Amazon_Mechanical_Turk_to_get_tasks_completed, accessed 17 December 2014.

Amazon (2014b), Worker Web Site FAQs, https://www.mturk.com/mturk/help?helpPage=worker#how_paid, accessed 17 December 2014.

Andreessen, M. (2011), "Why software is eating the world", *The Wall Street Journal*, 20 August, http://online.wsj.com/article/SB10001424053111903480904576512250915629460.html.

Atkinson, A.B. (1975), *Economics of Inequality*, Oxford University Press.

Aul, W.R. (1972), "Herman Hollerith: Data processing pioneer", *Think*, IBM's employee publication, November, pp. 22-24, www-03.ibm.com/ibm/history/exhibits/builders/builders_hollerith.html, accessed 25 May 2015.

Autor, D.H. and D. Dorn (2013), "The growth of low-skill service jobs and the polarization of the US labor market", *The American Economic Review*, Vol. 103, No. 5, pp. 1553-97, http://economics.mit.edu/files/1474, accessed 25 May 2015.

Bakhshi, H., A. Bravo-Biosca and J. Mateos-Garcia (2014), "Inside the datavores: Estimating the effect of data and online analytics on firm performance", Nesta, March, www.nesta.org.uk/sites/default/files/inside_the_datavores_technical_report.pdf, accessed 25 May 2015.

Bakhshi, H. and J. Mateos-Garcia (2012), "Rise of the datavores: How UK businesses analyse and use online data", Nesta, November, www.nesta.org.uk/sites/default/files/rise_of_the_datavores.pdf, accessed 13 May 2015.

Bamberger, K.A. and D.K. Mulligan (2011), "Privacy on the books and on the ground", *Stanford Law Review*, Vol. 63.

Barua, A., D. Mani, R. Mukherjee (2013), „Impacts of effective data on business innovation and growth", Chapter two of a three-part study, The University of Texas at Austin, www.businesswire.com/news/home/20100927005388/en/Sybase-University-Texas-Study-Reveals-Incremental-Improvement, accessed 20 May 2015.

Bekhuis, T. (2006), "Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy", Biomedical Digital Libraries 2006 Vol. 3, No. 2, *PubMed,* 3 April, http://dx.doi.org/10.1186/1742-5581-3-2.

Bertolucci, J. (2014), "Big data hot job: Data engineer", *Information Week,* http://www.informationweek.com/big-data/big-data-analytics/big-data-hot-job-data-engineer/d/d-id/1315579, accessed 25 May 2015.

Bloom, N. and J. Van Reenen (2007), "Measuring and explaining management practices across firms and countries", *The Quarterly Journal of Economics*, Volume 122, No. 4, November, pp. 1351-1408, http://stanford.edu/~nbloom/MeasuringManagement.pdf.

Brave, S. (2012), "We don't need more data scientists – just simpler ways to use big data", Gigaom, 22 December, https://gigaom.com/2012/12/22/we-dont-need-more-data-scientists-just-simpler-ways-to-use-big-data.

Bruner, J. (2013), "Defining the industrial Internet" O'Reilly Radar, 11 January, http://radar.oreilly.com/2013/01/defining-the-industrial-internet.html.

Bruno, L.C. (2014), "Plate, punch card, and instructions for Herman Hollerith's electric sorting and tabulating machine, ca. 1895", *American Memory*, Library of Congress, http://lcweb2.loc.gov/cgi-bin/query/r?ammem/mcc:@field(DOCID+@lit(mcc/023)), accessed 25 May 2015.

BLS (Bureau of Labor Statistics) (2015), *Occupational Employment Statistics*, November 2014, www.bls.gov/oes/home.htm, accessed 28 June 2015.

BLS (Bureau of Labor Statistics) (2014), *Occupational Outlook Handbook, 2014-15 Edition*, US Department of Labor, www.bls.gov/ooh, accessed 21 November 2014.

Brynjolfsson, E., L.M. Hitt and H.H. Kim (2011), "Strength in numbers: How does data-driven decision making affect firm performance?", Social Science Research Network (SSRN), 22 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

Brynjolfsson E. and L.M. Hitt (2000), "Beyond computation: Information technology, organizational transformation and business performance", *Journal of Economic Perspectives*, Vol. 14, No. 4, pp. 23-48.

Brynjolfsson, E. and A. McAfee (2014), *The Second Machine Age – Work, Progress and Prosperity in a Time of Brilliant Technologies*, W.W. Norton & Co., New York.

Brynjolfsson, E. and A. McAfee (2012a), *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity and Irreversibly Transforming Employment and the Economy*, January, Research Brief, http://ebusiness.mit.edu/research/Briefs/Brynjolfsson_McAfee_Race_Against_the_Machine.pdf, accessed 12 May 2014.

Brynjolfsson, E. and A. McAfee (2012b), "Big data: The management revolution", *Harvard Business Review*, October, https://hbr.org/2012/10/big-data-the-management-revolution/ar, accessed 12 May 2014.

Brynjolfsson, E and A. McAfee (2011), *Race Against the Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity and Irreversibly Transforming Employment and the Economy*, Digital Frontier Press, 17 October.

Carlisle, R. (2004), *Scientific American Inventions and Discoveries*, John Wiley & Sons, Inc., http://bioloskiblog.files.wordpress.com/2012/03/izumi-i-otkrica.pdf, accessed 12 May 2014.

c|net (2012), "Foxconn reportedly installing robots to replace workers", 13 November, www.cnet.com/news/foxconn-reportedly-installing-robots-to-replace-workers/.

Cingano, F. (2014), "Trends in income inequality and its impact on economic growth", *OECD Social, Employment and Migration Working Papers*, No. 163, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxrjncwxv6j-en.

Clearwater, A. and T.J. Hughes (2013), "In the beginning ... An early history of the privacy profession", *Ohio State Law Journal*, Vol. 74, p. 897, http://ssrn.com/abstract=2411814.

Coombes, A. (2013), "How to get investment advice for less online", *The Wall Street Journal,* 4 September, www.wsj.com/news/articles/SB10001424127887323477604578654101591319918.

Colvin, G. (2014), "In the future, will there be any work left for people to do?", *Fortune*, 2 June, http://fortune.com/2014/06/02/fortune-500-future/.

Cornia, G. and J. Court (2001), "Inequality, growth and poverty in the era of liberalization and globalization", UNU-WIDER, November, www.wider.unu.edu/publications/policy-briefs/en_GB/pb4/_files/78807311723331954/default/pb4.pdf.

Cowen, T. (2013), *Average is Over: Powering America Beyond the Age of the Great Stagnation*, Dutton Adult.

Cushing, E. (2013), "Amazon Mechanical Turk: The digital sweatshop", *UTNE*, January/February, www.utne.com/science-and-technology/amazon-mechanical-turk-zm0z13jfzlin.aspx, accessed 25 May 2015.

Cukier, K. and V. Mayer-Schönberger (2013), "The rise of big data – How it's changing the way we think about the world", *Foreign Affairs*, May/June.

Data Science Association (2012), "Code of conduct", www.datascienceassn.org/code-of-conduct.html, accessed 17 December 2014.

Davenport, T.H. (2014), "Cognitive technology – Replacing or augmenting knowledge workers?", CIO Journal, *The Wall Street Journal*, 18 June, http://blogs.wsj.com/cio/2014/06/18/cognitive-technology-replacing-or-augmenting-knowledge-workers/.

Davenport, T.H. and D.J. Patil (2012), "Data scientist: The sexiest job of the 21st century", *Harvard Business Review*, October, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1, accessed 25 May 2015.

Davenport, T. (2012), "Can you live without a data scientist?", *Harvard Business Review,* https://hbr.org/2012/09/can-you-live-without-a-data-sc, accessed 25 May 2015.

Davis, A.P. et al. (2013), "A CTD–Pfizer collaboration: Manual curation of 88 000 scientific articles text mined for drug-disease and drug-phenotype interactions", *Database*, http://dx.doi.org/10.1093/database/bat080, published online 28 November, 2013.

Dewan, S. and K. Kraemer (2000), "Information technology and productivity: Evidence from country-level data", *Management Science,* Vol. 46, No. 4, pp. 548-62.

Dunne, P.E. (2014), "Mechanical aids to computation and the development of algorithms", Lecture, Computer Science Intranet, University of Liverpool,

http://cgi.csc.liv.ac.uk/~ped/teachadmin/histsci/htmlform/lect4.html, accessed 25 May 2015.

Economist Intelligence Unit (2014), "Networked manufacturing: The digital future", *Economist Intelligence Unit* sponsored by Siemens, 7 July, www.economistinsights.com/technology-innovation/analysis/networked-manufacturing.

Economist Intelligence Unit (2012), "The deciding factor: Big data & decision making", commissioned by Capgemini, 4 June, www.capgemini.com/insights-and-resources/by-publication/the-deciding-factor-big-data-decision-making/.

Elliott, S. (2014), "Anticipating a luddite revival", in *Issues in Science and Technology*, Spring, p. 27-37; see also http://issues.org/30-3/stuart/, accessed 25 May 2015.

Essinger, J. (2004), *Jacquard's Web: How a Hand-loom Led to the Birth of the Information Age*, Oxford University Press.

Farrell, C. (2013), "It's time for a negative income tax", *Bloomberg Businessweek*, 8 August, www.businessweek.com/articles/2013-08-08/its-time-for-a-negative-income-tax.

Ford, M. (2009), *The lights in the tunnel: Automation, accelerating technology and the economy of the future*, Acculant Publishing, www.thelightsinthetunnel.com, accessed 25 May 2015.

Frey, C.B. and M. Osborne (2013), *The Future of Employment: How Susceptible Are Jobs to Computerisation?*, Oxford Martin School, University of Oxford.

Friedman, M. (1962), *Price Theory: A Provisional Text*, Kessinger Publishing (facsimile reprint, 2010).

Gino, F. and B.R. Staats (2012), "Samasource: Give work, not aid", Harvard Business School NOM Unit Case No. 912-011; UNC Kenan-Flagler Research Paper No. 2013-12, 22 February 2012, http://ssrn.com/abstract=2014220, accessed 26 May 2012.

Goldin, C. and L.F. Katz (2008), *The Race Between Education and Technology*, Belknap Press.

Goos, M., A. Manning and A. Salomons (2009), "Job polarization in Europe", *American Economic Review*: Papers & Proceedings Vol. 99, No. 2, pp. 58-63, http://dx.doi.org/10.1257/aer.99.2.58.

Gurulingappa, H. et al. (2013), "Automatic detection of adverse events to predict drug label changes using text and data mining techniques", Pharmacoepidem & Drug Safety, Volume 22, Issue 11, pp. 1189-94, Wiley Online Library, http://dx.doi.org/10.1002/pds.3493.

Harris, H., S. Murphy and M. Vaisman (2013), "Analyzing the analyzers: An introspective survey of data scientists and their work", *O'Reilly Strata,* June, www.oreilly.com/data/free/files/analyzing-the-analyzers.pdf, accessed 26 May 2015.

Herlyn, E. (2012), *Einkommensverteilungsbasierte Präferenz- und Koalitionsanalysen auf der Basis selbstähnlicher Equity-Lorenzkurven - Ein Beitrag zu Quantifizierung sozialer Nachhaltigkeit*, Wiesbaden.

Herlyn, E. et al. (2015), "Big data and analytics: What are the perspectives? Reflections on the OECD project on data-driven innovation", *OECD Digital Economy Papers*, OECD Publishing, Paris, forthcoming.

Herlyn, E., and F. J. Radermacher (2014), „Was kann Marketing für die Nachhaltigkeit tun? Eine Beobachterperspektive auf die Zukunft des Sustainable Marketing", in: *Sustainable Marketing Management*: *Grundlagen und Cases*, H. Meffert, P. Kenning and M. Kirchgeorg, (eds.), Springer Gabler Verlag.

Horton, J.J. (2011), "The condition of the Turking class: Are online employers fair and honest?", *Economics Letters*, Vol. 111, pp. 10-12, http://econpapers.repec.org/article/eeeecolet/v_3a111_3ay_3a2011_3ai_3a1_3ap_3a10-12.htm, accessed 26 May 2015.

Horton, J.J. and L.B. Chilton (2010), "The labor economics of paid crowdsourcing", Proceedings of the 11th ACM Conference on Electronic Commerce, pp. 209-18.

IBM (2014), "What is a data scientist", www-01.ibm.com/software/data/infosphere/data-scientist/, accessed 25 May 2015.

IBM (2013), "IBM Watson hard at work: New breakthroughs transform quality care for patients", News release, 8 February, www-03.ibm.com/press/us/en/pressrelease/40335.wss, accessed 25 May 2015.

ILO (2009), "Updating the International Standard Classification of Occupations (ISCO)", International Labour Organization, 19 March, www.ilo.org/public/english/bureau/stat/isco/docs/d334.doc, accessed 27 May 2015.

Insight Data Engineering (2014), "Insight Data Engineering Fellows Program", white paper, http://insightdataengineering.com/Insight_Data_Engineering_White_Paper.pdf, accessed 25 May 2015.

Ipeirotis, P. (2013), "Mechanical Turk account verification: Why Amazon disables so many accounts", Blog post, 28 June, www.behind-the-enemy-lines.com/2013/06/mechanical-turk-account-verification.html.

Ipeirotis, P. (2010), "Demographics of Mechanical Turk", working paper, hwww.researchgate.net/profile/Panos_Ipeirotis/publication/228140347_Demographics_of_Mechanical_Turk/links/00b7d51b0945c43fb5000000.pdf, accessed 7 December 2014.

$(ISC)^2$ (2011), *Annual Report 2010, Internet Information Systems Security Certification Consortium*, Palm Harbor, FL, www.isc2.org/uploadedFiles/(ISC)2_Public_Content/Annual_Reports/2010%20Annual%20Report.pdf.

Kan, M. (2013), "Foxconn to speed up 'robot army' deployment", *PCWorld*, 26 June, www.pcworld.com/article/2043026/foxconn-to-speed-up-robot-army-deployment-20000-robots-already-in-its-factories.html.

Kapitza, S. (2005), "Population blow-up and after: The demographic revolution and information society", Report to the Club of Rome and the Global Marshall Plan Initiative, Hamburg, www.clubofrome.org/?p=1785, accessed 26 May 2015.

Katz, D.M., M.J. Bommarito and J. Blackman (2014), "Predicting the behavior of the Supreme Court of the United States: A general approach", 21 July, http://ssrn.com/abstract=2463244 or http://dx.doi.org/10.2139/ssrn.2463244.

Kessler, A. (2004), *How We Got Here: A Silicon Valley and Wall Street Primer*, HarperBusiness, http://web.mit.edu/lin/Public/how-we-got-here.pdf, accessed 26 May 2015.

Keynes, J.M. (1930), "Economic Possibilities for our Grandchildren", W.W. Norton & Co., www.econ.yale.edu/smith/econ116a/keynes1.pdf, accessed 26 May 2015.

King, J. and R. Magoulas (2014), *2014 Data Science Salary Survey: Tools, Trends, What Pays (And What Doesn't) for Data Professionals*, O'Reilly Media, www.oreilly.com/data/free/2014-data-science-salary-survey.csp, accessed 26 May 2015.

King, J. and R. Magoulas (2013), *2013 Data Science Salary Survey: Tools, Trends, What Pays (And What Doesn't) for Data Professionals*, O'Reilly Media.

King, R.D. et al. (2004), "Functional genomic hypothesis generation and experimentation by a robot scientist", *Nature,* Vol. 427, No. 6971, pp. 247-252.

Kuonen, D. (2014), "A statistician's view on big data and data science", presentation at the Zurich Machine Learning and Data Science meetup in Zurich, Switzerland, 26 August, www.slideshare.net/kuonen/big-datadatascience-aug2014, accessed 26 May 2015.

Laney, D. (2012), "Defining and differentiating the role of the data scientist", Gartner blog, 25 March, http://blogs.gartner.com/doug-laney/defining-and-differentiating-the-role-of-the-data-scientist/.

Lee, T.B. (2015), "3D cameras are about to go mainstream: Here's why that's a big deal", Vox, 9 January, www.vox.com/2015/1/9/7520967/intel-realsense-3d-camera.

Levy, F. and R.J. Murnane (2013), "Dancing with robots: Human skills for computerized work", third way, 1 June, http://dusp.mit.edu/uis/publication/dancing-robots-human-skills-computerized-work.

Lodefalk, M. (2010), "Servicification of manufacturing - Evidence from Swedish firm and enterprise group level data", Working Papers No. 2010:3, Örebro University, School of Business, http://ideas.repec.org/p/hhs/oruesi/2010_003.html.

Lohr, S. (2009), "For today's graduate, just one word: Statistics", Inside Technology, *New York Times*, 5 August, www.nytimes.com/2009/08/06/technology/06stats.html.

Loukides, M. (2011), "What is data science? The future belongs to the companies and people that turn data into products", *O'Reilly Radar*, April, www.oreilly.com/data/free/what-is-data-science.csp.

Markoff, J. (2012), "Skilled work, without the worker", New York Times, 18 August, www.nytimes.com/2012/08/19/business/new-wave-of-adept-robots-is-changing-global-industry.html.

MGI (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey Global Institute, McKinsey & Company, June, www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

MacManu, R. (2009), "Report: Cloud-Based Email Cheapest Option for Most Companies", Readwrite, 6 January, www.readwriteweb.com/enterprise/2009/01/cloud-based-email-cheaper.php.

Marsan, C. (2012), "'Big data' creating big career opportunities for IT pros", Networkworld, 6 March, www.networkworld.com/article/2186578/data-center/-big-data--creating-big-career-opportunities-for-it-pros.html.

Mayer-Schönberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think,* John Murray, London.

MIT Technology Review (2010), "The truth about digital sweat shops", 12 January, www.technologyreview.com/view/417082/the-truth-about-digital-sweat-shops/.

Moore, G.E. (1965), "Cramming more components onto integrated circuits", *Electronics Magazine*, www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf, accessed 15 May 2015.

Mozur, P. and L. Luk (2012), "Hon Hai Hits Obstacles in Push to Use Robots", *The Wall Street Journal*, 11 December, www.wsj.com/articles/SB10001424127887324024004578172022369346936.

Myers, A. (2011), "Stanford team trains computer to evaluate breast cancer", Stanford Medicine News Center, 9 November, http://med.stanford.edu/news/all-news/2011/11/stanford-team-trains-computer-to-evaluate-breast-cancer.html.

NSF (2005), *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*, National Science Foundation, www.nsf.gov/pubs/2005/nsb0540/, accessed 15 May 2015.

O'Connor, S. (2013), "Amazon unpacked", *FT Magazine*, 8 February, www.ft.com/intl/cms/s/2/ed6a985c-70bd-11e2-85d0-00144feab49a.html#slide0, accessed 24 March 2015.

OECD (2015), *Digital Economy Outlook 2015*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264232440-en.

OECD (2014a), 2014 Ministerial Council Statement, OECD Publishing, Paris, 7 May, C/MIN(2014)15/FINAL, www.oecd.org/mcm/2014-ministerial-council-statement.htm, accessed 15 May 2015.

OECD (2014b), "ICT, jobs and skills: Proposals for a research agenda", OECD Publishing, Paris, 16 June, DSTI/ICCP/IIS(2014)2.

OECD (2014c), "Cloud computing: The concept, impacts and the role of government policy", *OECD Digital Economy Papers*, No. 240, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxzf4lcc7f5-en.

OECD (2014d), "Challenges and opportunities for innovation through technology: The convergence of technologies", OECD Publishing, Paris, 25 September, www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/stp(2013)15/final&doclanguage=en, accessed 26 May 2015.

OECD (2014e), "Inequality hurts economic growth, finds OECD research", OECD Newsroom, OECD Publishing, Paris, 9 December, www.oecd.org/newsroom/inequality-hurts-economic-growth.htm, accessed 15 May 2015.

OECD (2014f), *Education at a Glance 2014: OECD Indicators,* OECD Publishing, Paris, www.oecd.org/edu/eag.htm, accessed 28 June 2015.

OECD (2014g), "Skills and jobs in the Internet economy", *OECD Digital Economy Papers*, No. 242, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxvbrjm9bns-en.

OECD (2013a), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'big data'", *OECD Digital Economy Papers,* No. 222, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k47zw3fcp43-en.

OECD (2013b), *Supporting Investment in Knowledge Capital, Growth and Innovation*, OECD Publishing, Paris, www.oecd-ilibrary.org/industry-and-services/supporting-investment-in-knowledge-capital-growth-and-innovation_9789264193307-en, accessed 15 May 2015.

OECD (2013c), "Privacy Expert Group Report on the Review of the 1980 OECD Privacy Guidelines", OECD Publishing, Paris, http://dx.doi.org/10.1787/5k3xz5zmj2mx-en.

OECD (2013d), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264204256-en.

OECD (2012a), *Better Skills, Better Jobs, Better Lives: A Strategic Approach to Skills Policies*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264177338-en.

OECD (2012b), "ICT skills and employment: New competences and jobs for a greener and smarter economy", *OECD Digital Economy Papers*, No. 198, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k994f3prlr5-en.

OECD (2010), *OECD Information Technology Outlook 2010*, OECD Publishing, Paris, http://dx.doi.org/10.1787/it_outlook-2010-en.

OECD (2008), "Broadband and the economy: Ministerial background report, OECD ministerial meeting on the future of the Internet economy, Seoul", OECD Publishing, Paris, https://innovationpolicyplatform.org/document/broadband-and-economy-ministerial-background-report-oecd-ministerial-meeting-future, accessed 26 May 2015.

OECD (2005), "New perspectives on ICT Skills and Employment", OECD Publishing, Paris, www.oecd.org/internet/ieconomy/34769393.pdf, accessed 15 May 2015.

O'Reilly, T. (2014), "IoTH: The Internet of things and humans", *O'Reilly Radar*, 16 April, http://radar.oreilly.com/2014/04/ioth-the-internet-of-things-and-humans.html.

Patil, D.J. (2011), "Building data science teams", *O'Reilly Radar*, 16 September, http://radar.oreilly.com/2011/09/building-data-science-teams.html.

Piketty, T. (2014), *Capital in the Twenty-First Century*, Harvard University Press.

Provost, F. and T. Fawcett (2013), *Data Science for Business,* O'Reilly Media.

Radermacher, F.J. and B. Beyers (2007), *Welt mit Zukunft – Überleben im 21. Jahrhundert* [World with a future – Surviving in the 21st century], Murmann Verlag, Hamburg (2nd edition of "Welt mit Zukunft – Die Ökosoziale Perspektive, 2001).

Randers, J. (2012), *2052: A Global Forecast for the Next Forty Years*, Chelsea Green Publishing.

Reimsbach-Kounatze, C. (2015), "The proliferation of data and implications for official statistics and statistical agencies: A preliminary analysis", *OECD Digital Economy Papers*, No. 245, OECD Publishing, Paris, http://dx.doi.org/10.1787/5js7t9wqzvg8-en.

Rifkin, J. (2014), *The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism*, Palgrave Macmillan.

Rockwell, M. (2013), "Gold in the data, but a shortage of miners", *FCW*, 17 September, http://fcw.com/Articles/2013/09/17/big-data-analyst-shortage.aspx?p=1, accessed 21 May 2015.

Rogers, S. (2012), "What is a data scientist?", *The Guardian*, Data Blog, 2 March, www.theguardian.com/news/datablog/2012/mar/02/data-scientist.

Ross, J. et al. (2010), "Who are the crowdworkers? – Shifting demographics in Amazon Mechanical Turk", *CHI EA 2010*, pp. 2863-72, www.international.ucla.edu/media/files/SocialCode-2009-01.pdf?AspxAutoDetectCookieSupport=1, accessed 15 May 2015.

Royster, S. (2013), "Working with big data", Bureau of Labor Statistics, US Department of Labor, www.bls.gov/careeroutlook/2013/fall/art01.pdf, accessed 15 May 2015.

Solte, D. (2009), *Global Financial System in Balance – Crisis as Opportunity for a Sustainable Future,* Terra-Media, Berlin.

Spiegel Online (2014), "Apple-Zulieferer: Foxconn will Arbeiter durch Roboter ersetzen", Spiegel Online, 8 July, www.spiegel.de/netzwelt/gadgets/foxconn-will-arbeiter-durch-roboter-ersetzen-a-979819.html.

Stewart-Smith, H. (2012), "Foxconn chairman compares his workforce to 'animals'", *ZDnet*, 20 January, www.zdnet.com/blog/asia/foxconn-chairman-compares-his-workforce-to-animals/776.

Taleb, N.N. (2010), *The Black Swan: The Impact of the Highly Improbable*, Random House Trade Paperbacks.

Taleb, N.N. (2005), *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*, Random House Trade Paperbacks.

Tambe, P. (2014), "Big data investment, skills, and firm value", *Management Science*, http://ssrn.com/abstract=2294077, accessed 15 May 2015.

The Economist (2014), "Coming to an office near you", 18 January, www.economist.com/news/leaders/21594298-effect-todays-technology-tomorrows-jobs-will-be-immenseand-no-country-ready.

The Economist (2012), "High-frequency trading: The fast and the furious", 25 February, www.economist.com/node/21547988.

The Economist (2010), "Data, data everywhere", 25 February, www.economist.com/node/15557443.

Techlist (2014), "50,000 MTurk workers who have performed techlist HITs", http://techlist.com/mturk/global-mturk-worker-map.php, accessed 15 May 2015.

TNO (2013), "Thriving and surviving in a data-driven society", 24 September, http://publications.tno.nl/publication/34610048/xcv74S/TNO-2013-R11427.pdf.

Turing, A.M. (1950), "Computing machinery and intelligence", *Mind*, pp. 433-60.

Uddin, Z. (2012), "The dystopian digital sweatshop that makes the Internet run", *AlterNet*, 11 September, www.alternet.org/labor/dystopian-digital-sweatshop-makes-internet-run.

UK National Career Service (2014), "Data analyst-statistician", Job Profiles, https://nationalcareersservice.direct.gov.uk/advice/planning/jobprofiles/Pages/dataanalyst-statistician.aspx, accessed 18 December 2014.

UN Global Pulse (2013), "Anatomy of a Pulse Lab", 19 December, www.unglobalpulse.org/anatomy-pulse-lab, accessed 17 December 2014.

Upbin, B. (2013). "IBM's Watson Gets Its First Piece of Business in Healthcare", *Forbes*, 8 February, www.forbes.com/sites/bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/.

Violino, B. (2014), "The hottest jobs in IT: Training tomorrow's data scientists", *Forbes*, 26 June, www.forbes.com/sites/emc/2014/06/26/the-hottest-jobs-in-it-training-tomorrows-data-scientists/.

Voce, C. et al. (2009), "Should Your Email Live in the Cloud? An Infrastructure and Operations Analysis", Forrester, 5 January, www.forrester.com/rb/Research/should_email_live_in_cloud_infrastructure_and/q/id/42980/t/2.

Wang, S. and R.M. Summers (2012), "Machine learning and radiology", *Medical Image Analysis*, Vol. 16, No. 5, pp. 933-51, www.ncbi.nlm.nih.gov/pubmed/22465077, accessed 15 May 2015.

Wilkinson, R. and K. Pickett (2010), *The Spirit Level: Why Equality is Better for Everyone*. Penguin.

Wing, J.M. (2008), "Computational thinking and thinking about computing", Philosophical Transactions "A" of the Royal Society, www.cs.cmu.edu/afs/cs/usr/wing/www/talks/ct-and-tc-long.pdf, accessed 15 May 2015.

Wu, C.F.J. (1997), "Statistics = Data Science?", University of Michigan, http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf, accessed 15 May 2015.

Zittrain, J. (2009), "The Internet creates a new kind of sweatshop", *Newsweek*, 12 August, www.newsweek.com/internet-creates-new-kind-sweatshop-75751, accessed 15 May 2015.

Further reading

OECD (2014), 2014 OECD Ministerial Statement on Climate Change, OECD Publishing, Paris, www.oecd.org/mcm/MCM-2014-Statement-Climate-Change.pdf, accessed 15 May 2015.

OECD (2013), *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264204256-en.

OECD (2008), "Broadband and the economy", OECD Digital Economy Papers, www.oecd-ilibrary.org/science-and-technology/broadband-and-the-economy_230450810820, accessed 28 June 2015.

The Economist (2012), "High-frequency trading: The fast and the furious", 25 February, www.economist.com/node/21547988, accessed 15 May 2015.

Wing, J.M. (2008), "Computational thinking and thinking about computing", Philosophical Transactions "A" of the Royal Society, www.cs.cmu.edu/afs/cs/usr/wing/www/talks/ct-and-tc-long.pdf, accessed 15 May 2015.

Chapter 7

Promoting data-driven scientific research

This chapter summarises the recent evolution of science – mainly thanks to the advent of data analytics – towards a more open and data-driven enterprise. It examines how new and evolving opportunities for interconnecting and sharing have led to what could be called citizen science. A discussion follows on the various impacts of open access to science, research and innovation on the business and science communities and on citizens. There are examples of organisations involved in open data efforts, and an exploration of the challenges and opportunities presented by data sharing. The focus then shifts to policies and practices in the OECD area and beyond, with the emphasis on infrastructure for data sharing. With unrestricted access to publications and data, firms and individuals may use and reuse scientific outputs to produce new products and services – but do scientists and researchers have the incentives or indeed the skills to perform these tasks?

In the next 50 years, as the technologies of information and knowledge accelerate, the nature of the scientific process will change even more than it has in the last 400 years. (Kevin Kelly, Google TechTalks, 9 May 2006)

Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorisation of reality. (Boyd and Crawford, 2011)

Information and communication technologies (ICTs), new data storage infrastructure and large-scale computing are modifying the way science and research are conducted, disseminated and diffused. (See Chapter 3 on the enabling role of ICTs for data-driven innovation, DDI.) The term often used to describe this transformation of science into a more open and data-driven enterprise is *open science.* All main phases of the data value cycle introduced in Chapter 1 apply and impact this transformation, its scientific processes and outcomes:

- *Data creation and data collection* – ICTs allow the collection of large amounts of real-time data that can serve as the basis for scientific experiments and research, contributing to make science increasingly data-driven. It is now possible to collect, generate, access, use and reuse research and scientific material (articles and data sets but also images or digital lab records) at no or extremely low marginal cost, and speed the transfer of knowledge among researchers and across scientific fields, opening up new ways of collaborating and new research domains. In addition, ICTs are creating new opportunities to organise and publish the content of research projects, scientific publications and large data sets, so as to make it immediately available to other scientists and researchers as well as potential users in the business community and society in general.
- *Data analytics and software* – ICTs are not only modifying the way scientific material is generated, collected and stored, but also helping promote deeper analysis of data through new software and applications that allow a faster and more exhaustive use of data in science and research. New opportunities also arise thanks to text and data mining techniques (see Box 7.1). Some scientific disciplines have historically been at the forefront of data collection and exploitation, by means of complex scientific experiments (as in the case of physics, for example), and have always had a long tradition of data validation and preservation. However, in recent years – and again, thanks to the advent of the ICTs – less historically data-intensive scientific fields (such as the humanities for example) have become increasingly data-driven. Different scientific domains are becoming increasingly interconnected: data generated in one field of research may nowadays be treated with models and techniques traditionally belonging to other fields of research. Typical examples are social science data that today may be treated with algorithms and methodologies traditionally belonging to physics or computer sciences, due to the large scale of the data sets newly generated and collected. Even those disciplines with a long tradition in data collection and lab experiments are moving towards more sophisticated simulations that are validated thanks to the large amount of data digitally available.
- *Scientific output and decision making* – As science becomes increasingly data-driven, its social and economic values are enhanced. Just as business or government data are being used to create new goods and services, scientific data are enhancing the quality and output of scientific research. Open research data allow scientists in one field to exploit data in other fields, and reveal relationships or patterns that were not visible before. It allows them to formulate and test new hypotheses and improve the predictability of scientific models. The value of scientific data is not solely economic; data can help capacity building (education, network formation); public engagement (or public understanding of science and technology); knowledge and material sharing among research institutions; and policy and practice revision.

Box 7.1. **Opportunities and challenges arising from text and data mining (TDM)**

"Text and data mining" (TDM) refers to an ensemble of computer science techniques to analyse and extract knowledge and information from large digital data sets (i.e. big data), by looking for trends and patterns unnoticeable to human eyes. TDM is useful and increasingly used by researchers in all fields, from historians who scan historical documents and archives to medical experts who find common patterns in medical records. TDM is equally well-established in fields such as astronomy and genetics; its methods and techniques are widely used both in the public and the private sector. TDM algorithms investigate large-scale data sets containing not only figures and numbers but also other types of digital records, including text, images and audio files. TDM enables the use of common techniques and makes connections between unconnected fields of research. This represents a huge opportunity for the development of innovation.

The use of TDM has important repercussions in the academic community. With the growing amount of published (and unpublished) academic articles (an estimated 50 million as of 2010), it is becoming impossible for scientists and researchers to manually access, read and analyse publications. TDM provides the potential for accessing, scanning and analysing publications by means of machines.

Research on TDM techniques has also advanced considerably in recent years. The number of academic articles published on TDM since the beginning of the 1990s reveals that the United States has so far produced 46.6% of the publications dealing with TDM, followed by the United Kingdom (11.1%), Chinese Taipei (8.8%), Canada (5.7%) and China (4.6%). Continental Europe lags behind, according to some because of the legal obstacles European researchers face when trying to advance knowledge in this field.

In fact, the growth of open access and open data raises more generally the issue whether researchers using TDM should or should not face restrictions. The uncertainty created by the current legal framework regarding the scope of protection of works and databases is bound to create obstacles to TDM activities. Moreover, transaction costs would rise if researchers had to reconcile the terms and conditions of non-standard or non-interoperable licences.

Within the OECD area, Japan and the United Kingdom have modified the national copyright framework to allow TDM during the last decade. In 2009, Japan introduced an exception to its national copyright law to permit TDM, with the aim of boosting the Japanese digital economy. More recently, the United Kingdom has introduced a specific exception in the Copyright, Designs and Patent Act to allow TDM activities to take place without the rights holder's prior authorisation, under conditions stated in the law. This amendment will enter into force in October 2014. Together with the Japanese provision, the United Kingdom provision is a step towards facilitating scientific research that may in time be followed by other jurisdictions.

Source: European Commission (2014), *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group;* Sergey Filippov (2014), *Mapping Tech and Data Mining in Academic and Research Communities in Europe,* Lisbon Council, 16/2014.

7.1. The evolving scientific enterprise

The evolution of scientific and research processes also has an impact on research addressing global challenges. Global challenges by definition affect many individuals in a large number of countries and cannot typically be addressed or managed by a small team of researchers working in isolation. The data generated at global level that relate to issues of global concern, such as the environment and climate change or the ageing population and health, may be more powerfully exploited if properly interconnected and used by large

networks of scientists and researchers world-wide. This may avoid duplication of efforts and enhance co-ordination in science and research. Recent examples of large-scale research projects with a social challenge focus also include the Human Brain project sponsored by the European Commission[1] and the BRAIN Initiative in the United States.[2] These projects are examples of collective, international, multidisciplinary efforts that combine multiple scientific disciplines from biology, medicine, high-performance computing and robotics, with the aim of advancing understanding of the human brain for the benefit of the society. The potential of open science and open data efforts to advance research to address Alzheimer's disease and dementia has been recently highlighted by the OECD expert consultation "Unlocking Global Collaboration to Accelerate Innovation for Alzheimer's Disease and Dementia" (OECD, 2014; see also Chapter 8 of this volume).

Finally, open science and open data have the potential to strengthen relations between the scientific community and society. The landscape of science communication has changed, and scientists have now a broader range of mechanisms (e.g. through social networks, personal scientific blogs, videos, interviews and discussion forums) to communicate with citizens. These new scientific communication mechanisms can help build public trust in science. In short, the advent of open science has the potential to increase opportunities for the diffusion of research results among the scientific community and to society. In addition, as science is managed and produced in a more open and transparent manner, the basis for "trust" between science and society will be subject to greater scrutiny from citizens. For a similar discussion on the trust enhancing-capacities of open data for government, see Chapter 10.

Data-driven science and research

Data and measurement have always been fundamental to science. The advent of new instruments and methods of data-intensive exploration has prompted some to suggest the arrival of "data-intensive scientific discovery", which builds on the traditional uses of empirical description, theoretical models and simulation of complex phenomena (BIAC, 2011). This could have major implications for how discovery occurs in all scientific fields (Hey and Trefethen, 2003; Jirotka et al., 2006; Anderson, 2004; Bell, Hey and Szalay 2004). Big data science allows the development of scientific experiments as well as computer-based algorithmic simulations, even in those fields that traditionally were less data-intensive than others.

Data analytics tools (such as machine learning and pattern recognition techniques) are increasingly used by scientists to gain knowledge of phenomena and to test or validate models. Big data allow computer-based experiments and simulations even in those fields where traditional lab experiments were impossible or too difficult to organise. Some have even challenged the usefulness of models in an age of massive data sets, arguing that with sufficiently large sets, machines can detect complex patterns and relationships that are invisible to researchers (Anderson, 2008; Bollier, 2010). In addition, big data science and algorithmic-based experiments and research in themselves represent an opportunity for innovation and scientific discovery: fields such as computer or data science are currently exploiting big data as an opportunity to develop new and more efficient algorithms for data analytics, to be used by researchers active in different disciplines and fields (both in the public and private sector).

New instruments such as super colliders and telescopes, but also the Internet as a data collection tool, have been key to new developments in science, as they have changed the scale and granularity of the data being collected. The Digital Sky Survey, for example, which started in 2000, collected more data through its telescope in its first week than had

been amassed in the history of astronomy (*The Economist*, 2010), and the new SKA (square kilometre array) radio telescope could generate up to 1 petabyte of data every 20 seconds (EC, 2010). Furthermore, the increasing power of data analytics has made it possible to extract insights from these very large data sets reasonably quickly. In genetics for instance, DNA gene sequencing machines based on big data analytics can now read about 26 billion characters of the human genetic code in seconds. This goes hand in hand with the considerable fall in the cost of DNA sequencing over the past five years.

These new developments, affecting all scientific instruments across all scientific fields, indicate the potential for a new era of discovery and raise new issues for science policy. These issues range from the skills that scientists and researchers must master to the need for a framework for data repositories that adheres to international standards for the preservation of data; defines common storage protocols and metadata; protects the integrity of the data; establishes rules for different levels of access; and defines common rules that facilitate the combining of data sets and improve interoperability (OSTP, 2010).

Diversity of scientific data

Scientific research data vary enormously – in type and volume, as well as in use and long-term value (see Chapter 4 for a comprehensive description of the different features of data). Four types of research data in particular are important in research.

Observational data come from telescopes, satellites, sensor networks, surveys and other instruments that record historical information or one-time phenomena (such as astronomical data from the Sloan Digital Sky Survey, SDSS). This category also includes social science research, such as demographic surveys. In many cases these data cannot be replicated and should be retained.

Experimental data may be captured from high-throughput machines (such as accelerators), through clinical trials and biomedical and pharmaceutical testing, or through other controlled experiments. Preservation is particularly important for experimental data where it is not feasible or ethical to replicate data gathering. This includes some data dealing with human subjects and endangered species.

Computational data are generated from large-scale computational simulations. Although such data can be regenerated by rerunning the simulation, there are two reasons why computational data may need to be preserved over the medium term (three or more years). First, the data may be used as the basis for substantive and subsequent analysis, visualisation, or data mining. Second, time on a computer for additional computations may not be available within a sufficiently short delay. This is a common occurrence for very large-scale computations that run on supercomputers shared by the research community, such as those found at US Department of Energy national laboratories and National Science Foundation (NSF) centres.

Reference data sets are highly curated data that are often in high demand by multiple scientific communities. Such data are created for purposes ranging from mapping the human genome and documenting proteins to amassing longitudinal data on economic and social status. The Worldwide Protein Data Bank and Panel Study of Income Dynamics are such reference data sets. With all these data, there is often a need to preserve ancillary materials, such as calibrations of instruments, parameters of experiments, and lab notebooks. While most large research data collections are produced and used by researchers, they are also valuable for public policy. Policy makers' needs for information about climate, seismology, oceanography, clinical trials and social science research surveys, endangered

species, indigenous sites, archaeological sites and sensitive security matters go well beyond the demands of research and become a matter of urgent public priority.

Interaction of data and research materials

Access to research materials and tools are important for advancing research and innovation. As science becomes more data driven, the issue of access to research materials can interact with the issue of open data in complementary or conflicting ways. The first interaction is around the patenting of research materials (e.g. biological materials); insufficient disclosure of difficult (first order results). A second interaction concerns the portfolio patenting strategies of firms or universities. While strategic from a licensing and commercial point of view, such patent portfolios can be structured as de facto propriety databases that could impede open research data efforts in academia, as many researchers rely on (patented) research tools.

Collaborative research platforms and citizen science

As science becomes more data-driven and ICTs offer new possibilities to interconnect and share, international platforms to promote collaborative and networked research have emerged. The goal of these consortia is to promote not only data sharing, but also and more generally information sharing, to facilitate the creation of joint research projects and activities among teams of researchers globally. They are typically either discipline-specific or around broad themes, often related to social challenges such as health or climate change and the environment, and north-south co-operation. In some cases, these collaborative platforms go beyond the involvement of the research community alone and aim to engage business sector actors, the civil society, and individuals more broadly.

Examples of global platforms include the following. The Open Source Drug Discovery (OSDD) is an online platform for drug discovery that brings together scientists from the OECD area and developing countries to develop therapies for diseases endemic in developing countries, such as malaria, tuberculosis and Leshmaniasis.[3] Future Earth is a consortium active in global research on climate change and the environment, managed by the International Council for Science[4] and open to scientists active in all disciplines. DIYbio.org is a platform founded in 2008 with the goal of creating a community of biologists exchanging information and discussing research themes. Other types of collaborative platforms are not restricted to researchers and scientists, but have the goal to reach a broader community in civil society. For example, the BiOS[5] Initiative (Biological Innovation for Open Society) is an effort to promote bio-related innovation (in fields such as agriculture, biotech, water and agronomy) in disadvantaged communities.

Collaborative efforts in science and research reach beyond the research community to increasingly involve citizens and "amateur researchers" at different stages of scientific processes, from the collection of data to the solution of more complex scientific problems. The involvement of non-professional scientific communities in science and research efforts is often referred to as *citizen science*.

The participation of amateurs in scientific processes – the interaction with professional scientists – is not a new concept; it dates from the 18th century and generally related to data collection or observation, in particular in disciplines such as ornithology and astronomy. However, the improvement in communication capabilities, the emergence of mobile devices, the increase in storage capacity for the information collected, the possibility of transmission of not only text but images and sounds, and (certainly) the existence of greater public awareness have all led to the emergence of the citizen science phenomenon.

There are different aspects of citizen engagement. One is related to the degree of public participation, with respect to the kind of role the non-professional is playing. A second aspect is related to the role of citizens in the selection of research streams that will be publicly financed. Finally, citizen science has specific organisational characteristics; these relate to the development of networks of both professional and non-professional personnel through dedicated events, as well as the need for technical support from the scientific to non-scientific communities (Holocher-Ertl and Kieslinger, 2013).

It has been argued that citizen science serves as a means of achieving several different objectives (Riesch, Potter and Davies, 2013); for example, it allows the development of a more democratic environment in science by engaging amateurs as well as professionals in research and scientific efforts. There is a clear willingness on the part of civil society to be directly involved in the scientific process, not only as observers or data collectors, but also as practitioners, planners and evaluators. Society's participation in the process could even lead decision makers to opt for research priorities based on amateur scientist conclusions or to revoke decisions previously taken. Such was the case in the London District of Deptford: the UK Environment Agency revoked a scrapyard's licence after data that citizens collected on noise levels showed that the yard's operation violated noise limits (Gura, 2013). In addition, the involvement of citizens in scientific projects tends to have an educational value, both implicit and explicit. While in the majority of projects the informal learning aspect of adult citizens is addressed, schools are increasingly considered an important target for the introduction and promotion of citizen science. Teachers play a significant role in facilitating the deployment of experiments and transmitting the socio-scientific values of their contributions to the young audience. Citizen involvement in scientific efforts may additionally have positive implications for the development of a scientifically aware culture.

Several activities help promote the engagement of citizens in science. The project "Amateurs as Experts" was a three-year study of volunteer naturalists, biodiversity scientists and policy makers involved in the UK Biodiversity Action Plan process. The project began in October 2002 and lasted three years. The aim was to enrol new actors into the formal UK biodiversity policy process and have them gain experience carrying out social experiments and analysing and assessing their progress, results and problems. Through ethnographic research methods, the study tried to clarify the social and knowledge dynamics while also fostering patterns of interaction between the volunteer naturalists, scientists and policy makers. The project was a cross-disciplinary research study, involving sociologists, anthropologists (Institute for Environment, Philosophy and Public Policy [IEPPP], Lancaster University) and natural scientists (Natural History Museum, London). Its objective was to develop effective biodiversity protection policies, in the United Kingdom and beyond.[6]

Other examples of crowdsourcing for technical skills that can solve scientific problems are, for example, online platforms where solutions to scientific problems are requested from the public. Examples of this kind of website include Kaggle,[7] a web-based platform for predictive modelling and analytics: private companies and research teams publish unsolved problems related to specific data sets, and data scientists from all over the world compete to find the best solutions and highest-performing algorithms. The crowdsourcing approach relies on the fact that there are countless strategies to solve the problem, each with a different computational efficiency. The best strategy wins and receives the prize money advertised by the firm or the research team posting the problem online. It is estimated that Kaggle connects around 200 000 data scientists world-wide.

7.2. Impacts of open access to science, research and innovation

The impact of open science and open data may be measured in multiple ways. On the one hand, greater access to scientific inputs and outputs can improve the effectiveness and productivity of the scientific and research system, by reducing duplication costs in collecting, creating, transferring and reusing data and scientific material; by allowing more research from the same data; by multiplying opportunities for domestic and global participation in the research process; and by ensuring more possibilities for testing and validating scientific results.

On the other hand, increased access to research results (in the form of both publications and data) can not only foster spillovers to scientific systems, but also boost innovation systems more broadly. With unrestricted access to publications and data, firms and individuals may use and reuse scientific outputs to produce new products and services. (See Chapter 4 for a comprehensive discussion on the potential opportunities arising from data reuse.)

Several surveys report access difficulties for academics in the United States and Europe. For instance, according to CED (2012) 15% of US and Canadian scholars from all disciplines reported their level of access not to be satisfactory. Ware and Monkman (2008) found that only 66% of scientists in Europe and the Middle East reported to have good or excellent access (85% in the United States). There were even lower numbers outside those regions. Barriers to access to scientific material for researchers due to the high cost of subscriptions are also reported, by the survey of Rowlands and Nicholas (2005) and Sparks (2005).

Developing countries in particular may benefit from open access to scientific material. Chan, Kirsop and Arunachalam (2005) note figures from a World Health Organisation survey: in countries with an annual GNP per capita of less than USD 1 000, around 56% of medical institutions have no subscriptions to journals; in countries with GNP per capita of between USD 1 000 and USD 3 000, the percentage of medical institutions with no subscriptions was lower, but still as high as 34%. This is why initiatives are in place to improve developing countries' access to scientific material. For example, the Research4Life programme is a public-private partnership among three United Nations agencies, two universities and major commercial publishers that enables eligible libraries and their users to access peer-reviewed international scientific journals, books and databases for free or for a small fee (Royal Society, 2012). Certain access journals have been created in developing countries themselves, such as the *African Journal of Health Sciences*.

Sharing data has always been considered a crucial activity for scientific research and widely accepted by the scientific community (Fienberg, Martin and Straf, 1985). There is some evidence that, as with open access to scientific publications, sharing data can increase the citation rate of scientific papers (Piwowar, Day and Fridsma, 2007; Piwowar and Vision, 2012) and it fosters good scientific behaviour (Mooney, 2011). Sharing data allows the use and reuse of data from other researchers and individuals (Groves, 2009), would also protect against faulty behaviours and fraud in science and research, and may help improve data collection and management. For all these reasons, data-sharing practices are often regarded positively by the research community (Cragin et al., 2010).

Impacts on the scientific community

Data sharing allows not only verification of scientific results but also the re-analysis of data for different purposes from the ones originally conceived. This enhances the utilisation of data, promotes the competition of ideas and research (Gardner et al., 2003),

and fosters collaboration (Brase et al., 2009; Piwowar and Chapman, 2008; Murray et al., 2009). Data sharing also reduces duplication of effort from different researchers attempting to collect the same data sets (Kowalczyk and Shankar, 2010).

Lakhani et al. (2007) found that disclosing information relating to scientific problems to a large group of outside solvers is an effective means of solving those problems. In addition, disclosure of problem information facilitates problem solving at the boundary or outside the discloser's fields of expertise, thanks to the transfer of knowledge from one field to another. Williams (2010) researched material relating to efforts to decode the human genome. She found that articles dealing with openly available genome sequences led to 30% more articles than those focused on sequences protected by IPRs. The advantage in publications and commercialisations generated by open sequences was notable. Hardisty and Haaga (2008) conducted research on medical articles and found that open access articles were read twice as often by mental health practitioners. In addition, a practitioner's reading of the open access article was associated with their recommending a more cutting-edge treatment. As for scientific publications, data sharing is especially important for researchers in developing countries with fewer means to undertake expensive and time-consuming data collection efforts (Arzberger et al., 2004).

Impacts on the business community and on citizens

Scientists and academics are not the only groups that can potentially benefit from more open data. The demand from the business sector and individual citizens to access research results in the forms of data is significant. For example, the usage data from PubMedCentral show that 25% of the daily unique users are from universities, 17% are from companies, 40% are individual citizens and the rest are government or other categories (UNESCO, 2012). A recent study on R&D-intensive SMEs in Denmark (Houghton, Swan and Brown 2011) found that 48% of those SMEs consider research outcomes very important for their business activities, and more than two-thirds reported difficulties in accessing research material. Ware (2009) conducted a survey on UK small and medium-sized enterprises and found evidence that the equivalent of 10% to 20% of articles were not easily accessible to his survey respondents. Finally, it has been argued that making research data publicly available may promote the public understanding of science, evidence-based practices, and citizen science initiatives (Kowalczyk and Shankar, 2010).

Quantifying impact

Several studies have attempted to estimate the impact of greater access to data on the economy. A recent analysis of UK organisations (Royal Society, 2012; CEBR, 2012) estimated that data was worth approximately GBP 25 billion to UK private and public sector organisations in 2011. The estimates are the cumulative results of GBP 17.4 billion GDP gained in business efficiency, GBP 2.8 billion derived by business innovation, and GBP 4.8 billion derived by business creation.

In the United States, data released by the National Weather Service are estimated to contribute to the development of the private sector meteorology market in an amount corresponding to approximately USD 1.5 billion (Spiegler, 2007). In 2008, the NASA Landsat satellite imagery of the Earth's surface environment became freely available on the Internet. The usage of this database increased from 19 000 scenes per year (when scenes were sold for USD 600 each) to 2.1 million scenes per year. Leading Silicon Valley companies such as Google (in particular Google Earth) use these images, and the open release is estimated to have generated direct benefits of more than USD 100 million per year

to the US economy. According to a recent estimate of the US Open Data initiative (data.gov), open data has the potential to generate more than USD 3 trillion per year in additional value in sectors such as finance, consumer products, health, energy and education.[8]

The European Commission's open data initiatives are expected to generate a yearly income of EUR 140 billion (EC, 2012). In addition, the OECD (2015a) estimated that the public sector information (PSI) market for the OECD area could be around USD 500 billion plus an additional USD 200 billion if barriers to use were removed, skills enhanced, and data infrastructure improved.

The calculation of estimates for the economic value specifically of research data and the related contribution to economic development is more problematic. Available estimates include that of Houghton and Sheehan (2009), who analyse the effects of increasing accessibility to public sector research outputs in Australia, and estimate that increased accessibility generates a return of AUD 9 billion over 20 years. Houghton et al. (2010) estimate that the open access archiving mandate for US Federal Research Agencies over a transitional period of 30 years may be worth around USD 1.6 billion, and up to USD 1.75 billion if no embargo period is in place. Around USD 1 billion would benefit the US economy directly, and the remaining amount would translate into economic spillovers to other countries. These figures would be significantly higher than the estimated cost of implementing open access archiving. JISC (2014) conducted a study on the economic impact of three UK data centres (the Economic and Social Data Service, the Archaeology Data Centre and the British Atmospheric Data Centre) and estimated that each of them could bring a twofold to tenfold return on investment over 30 years.

Open data and the involvement of supranational entities

International organisations play a critical role in promoting co-ordination, at international level as well as in the adoption of standards and norms related to the interoperability of data-sharing platforms. They are also involved in the promotion of an open data culture among scientists and researchers. Examples of international organisations playing these roles are listed in Box 7.2. In addition, international governmental organisations (IGOs) such as the OECD, UNESCO, the EU and the World Bank have been active in recent years in promoting open science efforts of their respective member and, in some cases, non-member countries. The OECD has been active in developing guidelines and principles on open science-related themes, including access to public sector information (see Box 7.3, and Annex of Chapter 10 for more information) and in research data (see Box 7.4) or. At European level, the European Union has adopted and promoted open data efforts in the most recent Framework Programme for Research and Innovation, Horizon 2020 (Box 7.5).

Box 7.2. **Examples of organisations involved in open data efforts**

The *International Council for Science* (ICSU, at www.icsu.org), is a non-governmental organisation gathering members of national scientific bodies and international scientific unions worldwide, representing 140 countries. ICSU was founded in 1931 to promote international scientific activity. The organisation's current mission is to promote international science co-operation for the benefit of society. ICSU identifies and address major issues of importance to science and society; facilitates interaction among scientists across all disciplines and countries; provides independent advice to stimulate dialogue among the scientific community and governments, civil society and the private sector. The ICSU 2012-17 strategic plan has identified the following priorities: i) international research collaboration; ii) science for policy (making); and iii) the universality of science. ICSU has recently published its statement on Open Access Principles.[1]

Box 7.2. **Examples of organisations involved in open data efforts (cont.)**

The *Committee on Data for Science and Technology* (CODATA, at www.codata.org), is an interdisciplinary scientific committee of ICSU. CODATA works to improve the quality, reliability, management and accessibility of science and technology data. It also promotes awareness and cross-border co-operation of scientists. The committee was established in 1966 by ICSU to promote globally the compilation, evaluation and dissemination of reliable numerical scientific data. Legally independent from ICSU, CODATA has 23 members across different continents. Country membership often takes place through national research councils. CODATA activities include both technical discussion on standards and interoperability and policy-level discussion on data issues. The committee works on different aspects of data, from research data to social science data, and public sector information (PSI) including government data. CODATA is concerned with all types of data resulting from experimental measurements, observations and calculations in every field of science and technology, including the physical sciences, biology, geology, astronomy, engineering, environmental science and ecology. Special emphasis is placed on data management problems common to different disciplines, and to data used outside the field in which they were generated.

The *Research Data Alliance* (RDA, at rd-alliance.org) has the goal of promoting data sharing to accelerate data discovery, use, reuse, standards and harmonisation. RDA is organised into working groups and interest groups around different themes; the groups comprise experts from different countries and belonging to different communities (academia, the business sector, governmental agencies). RDA was created in 2013 by a core group of organisations: the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation. Individuals may also apply for membership; today, RDA counts around 1 600 members from more than 70 countries.

The *EMBL-EBI* (www.ebi.ac.uk) – The European Bioinformatics Institute (EBI) is part of the European Molecular Biology Laboratory (EMBL), a non-profit organisation and basic research institute funded by 20 member states in Europe, Israel, and one associate member (Australia). EBI is a major European laboratory for the life sciences; it provides freely available data from life science experiments in the field of molecular biology. EBI maintains the world's most comprehensive collection of freely available up-to-date molecular databases. Its services allow scientists to share data, perform complex queries, and analyse the results. Database users can generally work locally by downloading EBI data and software, and can access different resources through EBI web services. EBI serves millions of researchers world-wide who are active in multiple fields of the life sciences, from clinical biology to agri-food research. EBI also offers training programmes to maximise the benefits of data available in the life sciences to researchers in academia and the business sector. Some 20% of EBI users are engaged in industrial R&D, and EBI has developed an Industry Programme to collaborate specifically with firms active in bio-informatics. EBI addresses the specific needs of industry in other ways: from public-private partnerships to develop better and safer medicines for patients (the Innovative Medicines Initiative), to the provision of data infrastructure and services to SMEs, enabling bio-informatics spin-offs from EMBL and facilitating key pre-competitive research projects with industrial partners. EBI is located on the Wellcome Trust Genome Campus in the United Kingdom.

The *European Organization for Nuclear Research* (CERN, at web.cern.ch) is an international research laboratory containing the world's largest and most complex scientific instruments to study fundamental particles. CERN was founded in 1954 and is located on the Franco-Swiss border. It was one of Europe's first inter-country joint ventures, and it has now 21 member states. CERN actively supports open access efforts. In particular, since 1 January 2014, CERN hosts the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3). Supported by partners in 24 countries, SCOAP3 works in collaboration of over one thousand libraries, library consortia and research organisations to make available free of charge scientific articles in the field of high-energy physics. The consortium benefits from the support of funding agencies and has been established in co-operation with the publishing industry. As a result its efforts, articles are open access, the copyright stays with the author(s), and licence agreements allow text and data mining.

1. More information available at: www.icsu.org/general-assembly/news/ICSU%20Report%20on%20Open%20Access.pdf.

Box 7.3. **The OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information**

This OECD recommendation was developed by the OECD Committee for Information, Computer and Communication Policy and by its Working Party on the Information Economy. The recommendation was adopted by the OECD Council in 2008. The recommendation refers to the following items:

- *openness* – maximise the availability of public sector information for use and reuse, by taking into account limitations related to privacy and security, and in accordance to copyright
- *access and transparent conditions for reuse* – promote use and reuse by removing unnecessary restrictions. Improve access over the Internet and in electronic form
- *asset lists* – strengthen awareness of the kind of public sector information that is available for reuse
- *quality* – ensure the use of methodologies to enhance the quality and reliability of public sector information
- *integrity* – maximise the integrity and availability of information
- *new technologies and long-term preservation* – improve the usage of interoperable systems, and develop the necessary skills for preservation and access
- *copyright* – intellectual copyright should be respected
- *pricing* – price public sector information transparently when it is not provided for free
- *competition* – ensure that pricing strategies take into account considerations of unfair competition when both public and business users provide value added services
- *redress mechanisms* – provide appropriate transparent complaint and appeals processes
- *public-private partnerships* – promote public-private partnerships where appropriate and feasible
- *international access and use* – seek consistency in access regimes to facilitate cross-border use, improve interoperability, and support international co-operation and co-ordination
- *best practices* – share best practices and exchange information on implementation, the education of users, cost and pricing models, copyright handling, monitoring performance and compliance, and the wider impacts – on innovation, entrepreneurship and economic growth, as well as socially.

Source: Adapted from OECD, 2008.

Box 7.4. **The OECD principles and guidelines for access to research data from public funding**

In 2004, ministers of science and technology of OECD countries met in Paris and discussed the need for international guidelines on access to research data. At that meeting a *Declaration on Access to Research Data from Public Funding* was adopted by OECD countries. Following that meeting, the OECD's Committee for Scientific and Technological Policy (CSTP) launched a project to develop a set of principles and guidelines. Those that resulted from this project were approved by CSTP in October 2006 and then endorsed by the OECD Council. The principles can be summarised as follows:

- *Openness* – Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.
- *Flexibility* – Flexibility requires taking into account the rapid and often unpredictable changes in ICTs, the characteristics of different research fields, and the diversity of research systems, legal frameworks and cultures among member countries.
- *Transparency* – Information on research data and data-producing organisations, documentation on the data, and conditions attached to the use of data should be internationally available in a transparent way, ideally through the Internet.
- *Legal conformity* – Data access arrangements should respect the legal rights and legitimate interests of all stakeholders in the public enterprise. Restriction to access may be for reasons of national security; privacy and confidentiality; trade secrets and intellectual property rights; protection of rare, threatened or endangered species; or legal processes.
- *Protection of intellectual property* – Data access arrangements should consider the applicability of copyright and other intellectual property laws that may be relevant to publicly funded research databases (as in the case of public-private partnerships).
- *Formal responsibility* – Access arrangements should promote the development of rules and regulations dealing with the responsibilities of the various parties involved; should be developed in consultation with representatives of all parties affected; and should be responsive to factors such as the characteristics of the data and their potential value for research purpose. Data management plans and long-term sustainability should also be considered.
- *Professionalism* – Institutional arrangements for the management of research data should be based on the relevant professional standards and values embodied in the codes of conduct of the scientific communities involved.
- *Interoperability* – Access arrangements should consider the relevant international data documentation standards.
- *Quality* – The value and utility of data depend to a large extent on the quality of the data. Particular attention should be paid to ensuring compliance with explicit quality standards.
- *Security* – Attention should be devoted to supporting the use of techniques and instruments to guarantee the integrity and security of research data.
- *Efficiency* – One of the central goals of promoting data access and sharing is to improve the efficiency of publicly funded scientific research so as to avoid expensive and unnecessary duplication of effort. This also involves cost and benefit analysis to define data retention protocols, the engagement of data management specialist organisations, and the development of new reward structures for researchers and database producers.
- *Accountability* – The performance of data access arrangements should be subject to periodic evaluation by user groups, responsible institutions and research funding agencies.
- *Sustainability* – Due consideration should be given to the sustainability of access to publicly funded research data as a key element of the research infrastructure.

Source: OECD, 2007.

Box 7.5. **Open Data under Horizon 2020**

The new EU framework programme for research and innovation 2014-20 includes a pilot project on open research data. Researchers involved in projects participating in the pilot will be asked to make publicly available the data constituting the basis of the project research results; these can then be used by other researchers and projects, innovative industries and citizens. The researchers will also be asked to develop data management plans. Over 2014-15 the Open Research Data Pilot will receive around EUR 3 billion. The pilot project targets all key thematic areas of Horizon 2020 (future and emerging technologies, research infrastructure, leadership in enabling and industrial technologies, societal challenges, and science with and for society).

Researchers participating in the pilot have the possibility of opting out of it to protect intellectual property or personal data; for security concerns; or if the main objective of their research can be compromised by making data openly accessible. The pilot has the goal of providing a better understanding of what supporting infrastructure is needed, as well as the role of limiting factors such as security, privacy and data protection that could induce researchers to opt out. The Pilot also aims to contribute to a better understanding of the best mechanisms to define the right incentives for researchers to curate and share the research data they produce. The pilot will be closely monitored during the implementation phase of Horizon 2020.

Source: http://europa.eu/rapid/press-release_IP-13-1257_en.htm.

Data-sharing challenges and opportunities

An essential element for the usefulness of data-sharing efforts is the quality of the publicly released data. In many scientific communities there is as yet no standard data quality assessment protocol, as there is for scientific publications (Brase et al., 2009). As highlighted for example in Royal Society, 2012 and Dallmeier-Tiessen et al., 2012, data have little value if they do not meet minimum quality criteria. Data quality here implies being not only accessible (for example available on the Internet), but also intelligible, assessable, trustworthy and, of course, reusable. In this respect the development of detailed data-sharing information and metadata is essential for the further use of the same data by multiple teams of researchers. Additional challenges relate to privacy, or the definition of the ownership of the data itself. The OECD Global Science Forum has recently identified nine challenges related to data sharing (see Box 7.6).

As in the case of access to publications, data collection, curation and sharing vary by scientific discipline. Some fields have been traditionally more data-intensive than others, especially those making use of large-scale experiments managed by teams of hundreds of researchers (for instance the case with data generated through particle accelerators at CERN), or making use of machine-collected or -generated data. Researchers belonging to other scientific disciplines, notably in the social sciences and humanities, traditionally collect and built their own data sets, in some cases manually or by developing surveys and questionnaires. That makes this kind of the data set more tied to the individual researcher, and therefore less easily ready to be shared without proper curation, cleaning and metadata compilation (see Box 7.6, Challenge 4).

As highlighted by the OECD Global Science Forum (Box 7.6, Challenges 7 and 9), scientists and researchers do not have necessarily the incentives or the skills to perform these tasks, since proper curation and dissemination of data sets are costly and time-consuming, and can be even considered another type of scientific output (Uhlir, 2012). In addition, scientists and researchers traditionally compete to be the first to publish scientific results, and may not see the benefits of disclosing information on the data they want to use to produce as yet unpublished research outcomes.

Box 7.6. **Data sharing: Nine challenges identified by the OECD Global Science Forum**

The OECD Global Science Forum has recently identified a number of challenges related to data-driven and evidence-based research.

Challenge 1 – Massive amounts of digital data are being generated at unprecedented scale, partly thanks to the advent of ICTs. The reliability, statistical validity and "generalisability" of new forms of data are not yet fully understood.

Challenge 2 – While administrative, survey and census data are widely collected by national statistical agencies and government departments, micro-data records are much less available.

Challenge 3 – New forms of personal data, such as social networking data, are increasingly created and collected. The use of those data may generate risks to individuals' privacy.

Challenge 4 – Barriers to legal, cultural, language and proprietary rights of access hinder cross-national collaboration and international data exploitation, especially in the social sciences.

Challenge 5 – Global research agendas require increasingly interdisciplinary and international co-ordination.

Challenge 6 – Collaboration and experience sharing across countries in the development of comparable data resources is necessary to fully exploit the potential of data sets.

Challenge 7 – Researchers often lack the resources or the skills to make sure that the data they use, gather and produce are available for reuse.

Challenge 8 – National investments in skills and infrastructure related to data creation and curation are essential to avoid the risk of data loss or degradation.

Challenge 9 – Researchers need to have the right set of incentives to ensure effective data sharing.

Source: Adapted from OECD 2013.

A possible solution to the above-mentioned disincentives is data citation: researchers wishing to be acknowledged for their work could release data sets through mechanisms similar to the one already in place for citations of academic articles (Mooney and Newton, 2012; CODATA-ICSTI, 2013). Data citation is not, however, necessarily a standardised or widely accepted concept in the academic community. Some scientists see data citation as limiting citation to scientific articles. Funding agencies in some cases question the idea of recognising individuals as data authors, and traditional bibliometrics indicators are not yet taking into account non-article citations (Costas et al., 2013). There are in addition technical barriers restricting the development of data citation and related metrics. These include incompatibility – in machines and software, data file structures, data storage and management (Groves, 2010). Some organisations, such as DataCite (www.datacite.org; see Box 7.7), have been active in promoting conditions that will enable data citations, such as unique data object identifiers for data sets.

Box 7.7. **Organisations promoting data citations**

DataCite (www.datacite.org) is an international non-profit organisation established in London, United Kingdom since 2009. DataCite has the aims of promoting access to research data through the Internet; supporting open data archiving; and enabling verification of scientific results and the reuse of data for further studies. To facilitate data release, DataCite helps researchers with the unique identification and attribution of data sets for citation purposes, and supports journal publishers in establishing linkages between published articles and data sets. In addition, the organisation supports data centres by providing identifiers for data sets and defining workflows and standards for data publication.

ORCID (www.orcid.org) is a non-profit, community-driven organisation whose aims are to create and maintain a registry of unique researcher identifiers, and to link research activities and outputs on the basis of these identifiers. Research identifiers can be assigned not only to scientific articles but also to other forms of research output, including equipment, experiments, patents and data sets. Apart from the registry, ORCID provides application programme interfaces that support system-to-system communication and authentication. ORCID codes are available through open source licence.

Sources: DataCite website, at www.datacite.org; Orcid website, at www.orcid.org.

According to the EU FP7 research project Opportunities for Data Exchange, or ODE (Kotarski et al., 2012), there are certain unique features of data citation, owing to the particular properties of data sets. For instance, data sets may be of very different sizes and it is not always clear to what specific elements inside the data sets scholars are referring to – or, in the case of updates to the data sets, which version to cite. According to the ODE project, some of the good practices/challenges related to data citations are as follows:

- citation of the data set with identifier should be listed in a work's reference/bibliography, to enable tracking of citation metrics
- publishers need to provide authors and referees with guidance on data citation
- there is no clear agreement on the persistence or longevity requirements for data sets to be considered citable or cited
- there is lack of clarity and agreement on what authorship of data set means
- researchers need to promote awareness in their communities of the benefit of data citation and follow agreed data citation guidelines.

Other possible vehicles for publishing data sets are data journals, i.e. collections of scientific articles that specialise in publishing data papers. Data papers are articles with the primary purpose of describing data sets, rather than reporting on scientific investigation and analysis. Data papers contain facts about and descriptions of data. These papers aim to be a citable source of information on data that brings credit to the scholars who produced and described the database, disclosed detailed information on a data set, and brought the existence of the data to the attention of the scientific community (Chavan and Penev, 2011). Data journals may target broader scientific areas as well as specific domains, such as earth system science or geoscience.

7.3. Policies and practices: OECD countries and beyond

OECD and non-member countries are increasingly developing frameworks, guidelines and initiatives to encourage greater openness in science, both at national and supranational level (see Boxes 8.3 and 8.4 for recent OECD principles and guidelines related to access to research data or public sector information). Most of the respondent countries to the *OECD Science, Technology and Industry Outlook 2014* policy questionnaire highlighted recent changes in their policy framework for open science (OECD, 2014b). However, most open science efforts are broadly targeted, with the focus on developing the infrastructure necessary to promote sharing of research content and the collaboration platform of scientists and academics, or on requiring that the results of publicly funded research (mostly in the form of papers or publications) be made available on line, free of charge to the reader.

In other cases, open research data initiatives are nested into broader national open government or PSI strategies; they mainly have to do with releasing large sets of publicly collected data (such as weather or GIS data), which may or may not include research data. Some OECD countries and non-members, however, have adopted specific initiatives to promote open research data. Finally, a number of countries are modifying national copyright frameworks to promote greater use and reuse of data and research materials in science and research. This section will briefly review examples of infrastructure development for data sharing and open government initiatives, and then will cover more in detail those initiatives in OECD or non-member countries more specifically targeting open research data, the development of data analysis skills, and the involvement of business sector actors.

Developing the infrastructure for data sharing

Several countries are developing the infrastructure necessary to collect, store and disseminate research results (both articles and data). Examples of these initiatives include the creation of online repositories, databases, archives and digital libraries, as well as platforms containing information on R&D projects and researchers' CVs.

Argentina developed the SICyTAR (Sistema de Información de Ciencia y Tecnología Argentino) database containing information on the CVs, publications and affiliations of researchers. Colombia, Estonia, Greece, Mexico, South Africa, Korea and Poland have created national networks of repositories and digital libraries. Finland has launched an infrastructure roadmap to promote open science. People's Republic of China (hereafter 'China') has developed online platforms for data and publication archiving. Australia has been developing the eResearch infrastructure to help research organisations address the issues of data storing, accessing, analysing, modelling, sharing and manipulating. Germany also supports the development of open access libraries and archives, and has launched many initiatives on open access infrastructure for research outputs. In New Zealand, the Kiwi Research Information Service is the most comprehensive selection of publicly available research papers and related resources; materials include papers, conference materials and PhD theses, but also data sets. In France, many e-libraries and infrastructures for sharing research outputs have been developed, such as the National Hyper Articles Online Platform (HAL). In the United Kingdom, the E-infrastructure Leadership Council (ELC) advises the government on e-infrastructure aspects such as networks, data stores, computers, software and skills. It also advises BIS (business, innovation and skills) ministers on its antecedent implementation and

development. It works in partnership with stakeholders across the academic community, industry, government and society.

The European Commission has also been active in promoting the development of EU and member country repositories and platforms. In Poland, the new research data centre OCEAN has been announced, and will be operational in 2015 with the aim of providing the e-infrastructure for storage of open research data, as well as facilities and the expertise to undertake big data analytics. The target population of this centre is the entire Polish research community.

Including research data in broader open government/PSI agendas

Open government initiatives are high on the agenda of many OECD countries. The release of data collected by public administration may in some cases include research data or data on R&D activities. In addition, data provided through open government initiatives is contributing to open science, by giving to researchers and academics the opportunity to use, analyse and reuse those data, and potentially to advance in scientific discoveries and innovations.

The open government initiative in Canada is committed to making government data available and to promoting open data as a vehicle to further enhance the commercialisation of public research. In Sweden, the government supports the public release of data sets. In France, ETALAB co-ordinates the French open government data efforts, but it does not focus specifically on research data. In Norway, data.norge.no is the national portal for the storage and dissemination of publicly available data sets. A special Norwegian licence for public data was also developed in 2011 to facilitate reuse of data.norge data.

In the United Kingdom the Open Data Institute (ODI) has been created with the aim of promoting an open data culture to create economic, environment and social value. ODI helps to unlock data supply, generate demand and create and disseminate knowledge. ODI supports the UK Open Data agenda, and works in close co-operation with research centres and the business sector to advise how best to use and manage available data. Other open data initiatives in the United Kingdom focus on the personal data of consumers (for more informed choices), health and social welfare data, transport data, and education data. The US Open Government Initiative aims to improve access to federal data, including research data, to the public, through searchable, machine readable formats. Data are accessible through the Data.gov portal that also provides descriptions (metadata) on how to access and use the data sets. In Poland, access to public sector information is regulated by the Act of 2001. A new 2014 regulation defines the information assets as being places in dedicated repositories and the frequency of updates.

Targeting of research data

Some OECD and non-member countries have been recently implemented initiatives specifically to promote research data-sharing practices. The European Commission has in addition launched an open research data pilot project, under the new framework programme 2014-20, Horizon 2020 (see Box 7.5).

In Chile, since 2010, the National Commission of Technological Research (CONIYT) has been working on three main initiatives to generate access to research data from public funding: i) the development of a programme to manage research data and scientific information produced with public funding; ii) an analysis of the state of the art of

accessing and managing research data and scientific information; iii) the design of an institutional policy of management of research data and scientific information.

In China, the first pilot project of the Scientific Data Sharing programme was launched in 2001. At first restricted to meteorological data, the programme is now offering data-sharing opportunities in 24 scientific sectors, including the environment, agriculture, population and health. Major Chinese scientific institutions are users of the programme, which has promoted a data-sharing culture in Chinese research institutions.

The Finnish Research Data Initiative (preparatory phase 2009-11, project phase 2011-14) is part of the broader Finnish Open Science Roadmap, launched in 2014. The initiative was established to develop policy guidelines, improve ICT interoperability, establish collaboration platforms and provide service related to research data storage, long-term preservation, metadata, etc. The initiative was launched following completion of a survey of the status of research data management, which highlighted the need for action in these areas. The initiative identified a list of principles and guidelines related to research data, including:

- The principle of openness governing not only research data but also research methods such as computer models.
- Openness adhering to ethical principles and legal frameworks.
- Creative common licences are recommended to allow machine readability.
- Long-term preservation of data and infrastructure.
- Research data need to be documented and described.
- The need for each organisation to develop a data policy and related guidelines.
- Making references to data and methods developed by others is encouraged.
- Professional skills related to data need to be developed.
- Text and data mining should be allowed.
- Interoperability of international standards will be followed when possible.
- Researchers will be requested to consider the ownership of the data at the early stages of their research projects.
- Data management plans will be requested in every research plan.
- Copyright legislation will be amended to allow text and data mining for research purposes.

The United Kingdom is promoting the open release of data emerging from publicly funded research, as stated in the 2011 Research Strategy for Growth. Research Council guidelines on open access encourage the disclosure of research data and information on how to access it; however, data are not mandated to be open. A Data Capability Strategy has also been developed. The strategy focuses on three overarching aspects: i) human capital, a skilled workforce and data-confident citizens; ii) the tools and infrastructure necessary to store and analyse data; iii) data themselves as an enabler of research, innovation and economic development. In addition, a network of Administrative Data Research Centres (ADRN) has been created to promote research from interlinked administrative data. In 2012 the government agreed to introduce an exemption to the Freedom of Information Act to prevent the premature disclosure of research data and consequently having non-peer-reviewed data or results interpreted incorrectly by

journalists and other communities. In 2012, the government also established the Research Sector Transparency Board, which advises the government on how to increase access to research data. In addition, the UK copyright law framework has been recently amended (in 2014) to introduce targeted copyright exceptions that allow text and data mining for non-commercial research without specific authorisation.

In the United States, since the COMPETES Reauthorization Act of 2010, the US Office of Science and Technology Policy (OSTP) co-ordinates with US federal agencies to develop policies to promote public access to the results of federally funded research, including digital data. In 2011, two working groups on access to digital data resulting from government-funded research and access to scholarly publications were created. As a result of the efforts of these working groups, in February 2013 the OSTP issued a memorandum to federal government science agencies requesting the development of plans for increasing public access to the results of federally funded research, in particular to scientific publications and digital data. The OSTP memorandum establishes the following principles for agency policies:

- maximise free of charge general public access to digitally formatted scientific data collected with the support of federal funds, while protecting confidentiality and personal privacy, recognising intellectual property rights and preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden.
- make sure that researchers develop data management plans, including long-term preservation, or explain why long-term preservation and access are not possible
- allow the inclusion of appropriate costs for data management and access in proposals for federal funding of scientific research
- ensure evaluation of the data management plans submitted
- promote the storage of data in publicly accessible database repositories
- promote co-operation with the private sector, including by means of public-private partnerships to improve data access and compatibility
- develop mechanisms for data attribution and identification, to acknowledge researchers' open data efforts
- Support training, education, and workforce development related to scientific data management, analysis, storage and preservation.

In addition, individual American research institutions have developed open data policies. The US National Institutes of Health (NIH) developed a data-sharing policy as early as 2003 to encourage NIH-funded researchers to share scientific data sets. The policy requires of applicants requesting USD 500 000 or more of funding to include a data-sharing plan in the grant application procedure, or to justify why data sharing is not possible. Data-sharing plans should include a description of whether and how data will be made available, including how to account for protection of privacy, confidentiality, security and intellectual property rights; a description of the data to be shared; the timeline of sharing; data formats; procedures related to data-sharing agreements; and limitations on the use of data. The policy requires that data be shared no later than the acceptance for publication of the main findings from the final data set. Other NIH data-sharing policies are specifically developed to target different types of scientific data, collected during different projects researching different aspects of medicine, health and biological research.[9]

Investing in skills

In order to promote a transition towards data-driven research and innovation, a number of countries are investing in the skills necessary for data analytics. Data science training can be promoted at the postgraduate level, in order to provide PhD students or researchers with the skills necessary to make more extensive use of data analytics, or to promote the development of data management plans. Such training can also be promoted at the undergraduate level with the creation of new or adapted university curricula focused on science, technology, engineering and mathematics (STEM) skills, with the aim of delivering data scientist or data engineer degrees (see Chapter 6 on skills implications of DDI).

The European Commission, for example, estimates that Europe alone will face a shortage of up to 900 000 ICT professionals by 2020, due to a severe skill mismatch. To overcome this problem, in March 2013 the European Commission launched the Grand Coalition for Digital Jobs, a multi-stakeholder partnership to promote collaboration among business, education providers and public and private actors, to attract the young to ICT curricula and retrain unemployed people.

Countries severely hit by the financial crisis, such as Portugal, see the re-skilling or training of personnel in data-intensive areas – such as big data, data management and business analytics – as an opportunity to reduce the high domestic unemployment rates. Other countries are promoting the development of skills or ad hoc training in the broader context of open data and PSI agendas. The Italian Agency for Digital Italy has developed national guidelines for the exploitation of open data sets, bringing awareness of open access to data to academics, students, researchers and citizens. In Finland, in 2013, a working group established by the Finnish Ministry of Education and Culture produced a Data Management Guide that covers various aspects of that responsibility. The guide contains a checklist to assist researchers in data management planning, and provides information on available related services.

Other countries are developing specific training centres or higher education programmes. Poland has established a research data centre to provide extensive training curricula for big and open data management, and analysis to support interdisciplinary research in different scientific areas. In the United Kingdom, many initiatives are devoted to providing skills training in numerical data subjects, to teach students and academics how to use the big data sets emerging from open data efforts. The UK Engineering and Physical Sciences Research Council has announced a number of new Centres for Doctoral Training focusing on big data to be developed in several universities, including the University of Nottingham, the University of Edinburgh and the University of Oxford. In addition, the UK Economic and Social Research Council, the Higher Education Funding Council for England, and the Nuffield Foundation are supporting delivery of data science undergraduate programmes in 15 universities across the country, to promote quantitative social science training.

In the United States, the 2013 Office of Science and Technology Policy memorandum (see previous section) encourages the support of federal science agencies for training, education and workforce development related to scientific data management, analysis, storage, preservation, and stewardship. In addition, individual institutes have developed ad hoc skills policies. For example, the National Institutes of Health focuses on skills development through the initiative Big Data to Knowledge. The goal is to develop teams of researchers skilled in the science of big data and to increase the level of competencies in data usage and analysis across the biomedical research workforce. Governments also

focus on helping the ICT sector to better understand labour market shortages. Canada's Sectoral Initiatives Program, for example, includes training programmes to address industry skill mismatches by better aligning the skills of ICT and data specialists with the needs of employers (see Chapter 6).

The role of the business sector

Business sector actors are involved in data-intensive scientific activities, in several ways. Business organisations can be the actors providing the infrastructure to store, maintain and curate large-scale data sets and related services, such as the provision of processed data and of specific database extractions that may be relevant for the research community. With respect to open access to scientific publications, private scientific publishers have traditionally been the ones offering the services of article peer review (to guarantee quality) and digital and paper publication. Given the increasing data intensity of science, the range of services offered by business organisations to the research community are likely to diversify and expand. Some new services have already emerged, like those based on software and applications to catalogue and organise scientific articles and libraries or the data sets containing bibliometric records (see Box 7.8).

Box 7.8. **Start-ups for open data: The case of Figshare**

Figshare is an online digital repository of research data that includes figures, images and videos. Figshare was launched in 2011 by a PhD student in London. Figshare users can make their research outputs available to other researchers or users in a "citable, sharable and discoverable manner". This means that they can easily share data, search for data sets, and get credit for the data sets they upload on the website (through data set citations). Figshare allows users to upload any file format.

Figshare has recently established partnerships with other open science business actors, such as the open access publishing company PLOS ONE, the Nature Publishing Group, Taylor and Francis and F1000. These partnerships allow authors to directly upload data sets linked to papers online and ORCID (see Box 7.2), a service promoting data set citations. In addition, Figshare tracks the number of downloads of research materials, and it is often used as a source of alternative metrics. All files uploaded on Figshare are released under a creative commons licence.

Figshare stores more than 1.5 million files. From its original location in the United Kingdom it has today expanded to the United States and Romania. Users can sign in and upload or download content on Figshare for free. The company does however charge for premium services (such as larger private online storage space or private collaborative spaces) to individual researchers, and for services offered to publishers. In addition, it recently launched Figshare for Institutions, a service that is explicitly designed for research institutions around the globe.

Source: The Figshare website, at www.figshare.com.

The business community is also involved in open science and open data, as private firms can be the beneficiaries of open access publications and data that they use to develop new products and services and promote innovation more generally. In addition to entirely privately funded initiatives, joint public-private initiatives have begun to emerge for the delivery of services related to open science and open data. In other cases, the business community may hold and collect culturally and scientifically valuable material. There may be a

public interest in curating and preserving that material, and public-private partnerships can be developed to support preservation. This is the case, for instance, with the US National Film Preservation Board and the National Recording Preservation Board, which each year select items (often belonging to private organisations) to be added to their national registries (Blue Ribbon Task Force, 2010). Other examples of public-private partnerships involved in open science have recently been developed in Finland (Box 7.9).

Box 7.9. **Public-private partnerships for open science: The Finnish SHOK and DIGILE**

The Strategic Centres for Science, Technology and Innovation (SHOK in Finnish), established in Finland, are new public-private partnerships aiming to speed up innovation processes through renewal of clusters and the development of radical innovations. SHOK centres develop and apply new methods for co-operation, co-creation and interaction.

One such strategic centre is directly involved in open science: DIGILE, whose mission is to create digital business ecosystems that enable new global growth business for DIGILE's owners and partners. DIGILE aims not only to bring together R&D communities, but also to make sure that the results of scientific processes are understood, applied and adopted by companies. There are over 30 partners, including companies, research institutes and universities. DIGILE's strategy for 2015 focuses on data sharing, management and reusing as well as innovative data-intensive business models and services.

Source: DIGILE's website at www.digile.fi, Finland country note in OECD (2015b), and Myllymäki, P. (2013), "Data to Intelligence (D2I) Research Programme on Intelligent Data-Driven Services" presentation on www.digile.fi.

In addition to public science and innovation actors, private non-profit organisations and foundations may play a significant role in developing, raising awareness of and encouraging an open science culture. They may not only fund open access research and introduce requirements in grant agreements, but also develop and facilitate the creation of networks of stakeholders worldwide.

Open research data and IP protection

The expansion of open access policies to publicly funded research data raises a number of legal and policy issues that are often distinct from those concerning the publication of scientific articles and monographs. Since open access to research data – unlike publications – is a relatively new policy objective, less attention has been paid to the specific features of research data.[10]

Internationally, the protection afforded to databases (as collections of data or other elements) is established – or confirmed – by both Art. 10(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) and in the almost identical Art. 5 of the WIPO Copyright Treaty (WCT). According to the former:

> *Compilations of data or other material, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creation shall be protected as such... (Art. 10(2) of the TRIPS Agreements)*

Databases are a particular subject matter that is protected by copyright under certain circumstances, but that in some areas – namely within the European Union, Japan and South Korea – is also protected by a so-called sui generis database right (SGDR). This additional layer of protection is found in some countries and is afforded to databases regardless of the intellectual creation (i.e. "selection or arrangement") that may or may not be present. What is protected instead is the investment in making the database, i.e. in the obtaining, verification or presentation of the data. This type of right is typical, for example, of the EU Database Directive. It should be borne in mind that while the original protection afforded to databases focuses on the arrangement or selection without extending to the content of the database, the SGDR offers protection against the copy of substantial parts of the database – that is to say it extends, at least to some extent, to the data themselves.

With respect to research data in particular, the complexity of their rights status in Europe and other jurisdictions arguably has the potential to adversely affect the reuse opportunities of the collections of scientific data, given the difficulty – both for research institutions making the database available and for prospective reusers – in determining each time whether a certain database is covered by a sui generis right and in which measure reutilisation and extraction can take place freely. It is uncertain whether the use of compilations or databases for purposes of research and private study in general, and text and data mining in particular, is covered by any relevant exception on copyright or the database right. The use of Creative Commons licences 4.0 may alleviate the uncertainty, by clearly stating what can and cannot be done with the licensed material.

Another legal issue that comes into play in the context of open data, but that is less sensitive in the case of open access to scientific publications, is privacy and personal data protection. Data gathered in the course of research often contain personal data (e.g. medical records), and so opening such data has to respect the rights of data subjects (Lane et al., 2014). This does not mean that the data cannot be opened, but it does calls for implementing protective procedures (see Chapter 5). One of them is anonymisation, which may lead to the inapplicability of the whole personal data protection regime. In some cases however, research results depend on personal data and personal characteristics, and a complete anonymisation is therefore not always possible. Additionally, not all anonymisation techniques are effective (Narayanan and Shmatikov 2008).

Hence, the open science movement faces more challenges with regard to data than to scientific publications. While sui generis rights issues are to a large extent addressed in model licences (Creative Commons, Open Database Licenses), there are still no standardised procedures to follow with regard to privacy. Access and reuse of PSI also remain evolving subjects.

7.4. Key findings and policy conclusions

The scientific enterprise is evolving and becoming more data-intensive. Recent advances in technology are radically changing the way in which data are collected, stored and used. Data are collected and generated more quickly and in larger volumes than ever before. While not entirely novel – some scientific fields have always relied on large data sets – the increase in existing data as a source of inference for scientific evidence is spreading and deepening across scientific fields, including the social sciences. Another characteristic is the role of machines and algorithms to make sense of the large amounts

of data. Consequently, science requires continued investment in infrastructure, both hard and soft (e.g. data science skills).

However, several barriers to data sharing still remain. Although there is a clear potential to improve science and innovation systems, barriers still remain with respect to data-sharing efforts. Some barriers are of a technical nature, such as issues related to storage, the technical infrastructure to allow data sharing, interoperability and standards. Other types of barriers are related to the lack of an open data culture or the disincentives that researchers and scientists face with respect to the disclosure of data sets, especially relative to research at the pre-publication stage. This raises the question of the "optimal" level of openness to boost research and innovation without discouraging data collection from individual researchers.

Open science and open data are hot topics in many OECD countries and beyond. Many OECD countries and non-members have recently adopted and developed initiatives to promote greater openness and sharing of publicly funded research outputs, in the form of both articles and data. As in the case of policies to promote access to publications, the policy measures to promote open data may be developed and adopted by diverse sets of actors at both national and sub-national levels, as well as at the institutional level (universities, public research institutes). Policy measures may include efforts and initiatives such as mandatory rules, incentive mechanisms or enablers:

- *Mandatory rules* are often implemented in the form of requirements in research grant agreements, or in some cases defined in national strategies or institutional policy frameworks.
- *Incentive mechanisms* may take the form of financial incentives to cover the release of data sets. They may also be in the form of proper acknowledgment of open data efforts of researchers and academics, for instance in the form of data set citations or career advancement mechanisms partly based on metrics that take into account open science or data-sharing initiatives.
- *Enablers* are for example the infrastructure developed to share data; initiatives undertaken to develop an open science and open data culture; amendments to the legal framework to make it increasingly open science-friendly; or development of the skills necessary for researchers to share and reuse the research outputs produced by others.

(Cross-)subsidising the production of scientific knowledge requires picking winners (users or applications) by assessing (social) demand for a public good such as scientific knowledge based on the (social) value it creates. Governments can support the production of public goods i) by directly producing these goods, or ii) by supporting private firms' production of public and social goods through (e.g.) research grants, procurement programmes, contracted research and tax incentives. All these strategies raise a number of issues, including but not limited to difficulties in picking winners and losers, and the fact that resources are limited. Open access regimes can be a more efficient and politically attractive "indirect intervention" to support the production of public and social goods (see Chapter 4). As Frischmann (2012) highlights, "commons management is not a direct subsidy to … users who produce public or social goods, but it effectively creates cross-subsidies and eliminates the need to rely on either the market or the government to 'pick winners' – that is, to prioritise or rank … users worthy of access and support".

There is *tension between open research data and IPRs,* and a balance must be struck between efforts to promote open data in science and efforts to promote commercialisation

of public research, especially in the case of public-private partnerships involving companies. The tension can however be lessened by policies that clarify IP ownership and promote non-exclusive licensing possibilities, as well as by greater IP awareness among researchers, including of copyright and databases.

There is an *increasing need to develop skills* related to data curation, cleaning and preservation, both in the research community and in society in general. A number of countries have begun to address the shortage of data management skills by requiring researchers to develop data management plans in grant agreements or by developing training programmes or new academic curricula. There is a general need to understand the demand for those skills and the type of skills currently lacking in the research community and beyond to fully reap the benefits of data sharing.

Coherent guidelines are needed to promote better access to data across the economy. Many of the barriers to data access in science are common to those encountered in other domains, including open government data (see Chapter 10). Existing frameworks that promote better access to data, some of which are sector specific, may need to be reviewed and eventually consolidated to foster coherence among public policies related to data access, linkage and reuse. This would also include the OECD Council Recommendations promoting better access to data, including in particular the OECD (2008) *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* of 30 April 2008, and the OECD (2006) *Recommendation of the Council concerning Access to Research Data from Public Funding* of 14 December 2006, both of which are currently under review.

Notes

1 More information is available at: www.humanbrainproject.eu.

2 More information is available at: www.nih.gov/science/brain/.

3 More information is available at: www.osdd.net.

4 More information is available at: www.icsu.org/future-earth.

5 More information is available at: www.bios.net.

6 More information is available at: www.lancaster.ac.uk/fss/projects/ieppp/amateurs/.

7 Kaggle has partnered with large organisations including NASA and Deloitte, see www.kaggle.com.

8 An ongoing study of the US GovLab Academy at New York University – an online community that uses technology and innovation to solve public domain problems – is attempting to understand how US companies use open government data, through the Open Data 500 project. The project is analysing US-based companies (including international companies with a major presence in the United States) using open government data, a critical resource for their business. Most of the companies in the study belong to the technology, financial and business/legal services industries. According to the study, the most widely used data originates from the Department of Commerce, followed by the Department of Health and Human Services and the Securities and Exchange Commission.

9 See www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html; http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm; and http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm.

10 This section is derived from a background paper prepared for the OECD by Lucie Guibault and Thomas Margoni, "Legal aspects of open science and open data", 2014.

References

Anderson, C. (2008), "The end of theory: The data deluge makes the scientific method obsolete", *Wired*, 23 June, www.wired.com/science/discoveries/magazine/16-07/pb_theory/, accessed 15 April 2015.

Anderson, W.L. (2004), "Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data", *Data Science Journal*, 30 December, pp. 191-202, www.jstage.jst.go.jp/article/dsj/3/0/191/_pdf.

Arzberger, P. et al. (2004), "Promoting access to public research data for scientific, economic and social development", *Science*, 3 November, pp. 1777-78.

Bell, G., T. Hey and A. Szalay (2009), "Beyond the data deluge", *Science*, 6 March, pp. 1297-98, www.cloudinnovation.com.au/Bell_Hey_Szalay_Science_March_2009.pdf, accessed 15 April 2015.

BIAC (2011), "BIAC thought starter: A strategic vision for OECD work on science, technology and industry", Business and Industry Advisory Committee to the OECD, 12 October, http://biac.org/wp-content/uploads/2014/05/02-FINAL-11-10-_THOUGHT_-STARTER-_-VISION_-PAPER.pdf, accessed 6 May 2015.

Blue Ribbon Task Force (2010), "Sustainable economics for a digital planet: Ensuring long-term access to digital information", Blue Ribbon Task Force on Sustainable Digital Preservation and Access, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf, accessed 15 April 2015.

Bollier, D. (2010), *The Promise and Peril of Big Data*, Aspen Institute, Washington, DC.

Brase, J. et al. (2009), "Approach for a joint global registration agency for research data", *Information Services and Use*, Vol. 29, No. 1, pp. 13-27.

CEBR (2012), "Data equity: Unlocking the value of big data", Centre for Economics and Business Research Ltd, London, www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf, accessed 15 April 2015.

Chan, L., B. Kirsop and S. Arunachalam (2005), "Open access archiving: The fast track to building research capacity in developing countries", *SciDevNet*, November.

Chavan V., Penev L., (2011), "The data paper: a mechanism to incentivize data publishing in biodiversity science", *BMC Bioinformatics*, **12**(Suppl 15):S2, doi:10.1186/1471-2105-12-S15-S2.

CODATA-ICSTI (Committee on Data for Science and Technology – International Council for Scientific and Technical Information) (2013), "Out of cite, out of mind: The current state of practice, policy and technology for the citation of data", *Data Science Journal*, 13 September, www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article.

Committee for Economic Development (CED), (2012), *The Future of Taxpayer-Funded Research: Who Will Control Access to the Results?*, Washington D. C.

Costas, R. et al. (2013), "The value of research data: Metrics for datasets from a cultural and technical point of view", Knowledge Exchange report, www.knowledge-exchange.info/datametrics, accessed 15 April 2015.

Cragin, M.H. et al. (2010), "Data sharing, small science and institutional repositories", *Philosophical Transactions of the Royal Society A*, Vol. 368, No. 1926, pp. 4023-38.

Dallmeier-Tiessen, S. et al. (2012), "Compilation of results on drivers and barriers and new opportunities". Retrieved from www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/08/ODE-CompilationResultsDriversBarriersNewOpportunities1.pdf , accessed 15 April 2015.

EC (2010), "Riding the Wave: How Europe can gain from the rising tide of scientific data", Final report by the High-level Expert Group on Scientific Data, European Union, October, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf, accessed 15 April 2015.

EC (2012), "Public sector information – Raw data for new services and products", http://europa-eu-audience.typepad.com/en/2010/08/ec-public-sector-information-raw-data-for-new-services-and-products.html, accessed 15 April 2015.

Fienberg, S.E., M.E. Martin and M.L. Straf (1985), *Sharing Research Data*, National Academies Press, Washington, DC.

Filippov, S. (2014), "Mapping tech and data mining in academic and research communities in Europe", *Lisbon Council Newsletter*, Issue 16/2014.

Frischmann, B.M. (2012), *Infrastructure: The Social Value of Shared Resources*, Oxford University Press.

Gardner, D. et al. (2003), "Towards effective and rewarding data sharing", *Neuroinformatics*, Vol. 1, No. 3, pp. 289-95.

Groves, T. (2010), "The wider concept of data sharing: View from the BMJ", *Biostatistics*, Vol. 11, No. 3, pp. 391-92.

Groves, T. (2009), "Managing research data for future use", *The BMJ*, 338, b358

Gura, T. (2013), "Citizen science: Amateur experts", *Nature*, No. 496, pp. 259-261, naturejobs.com, http://dx.doi.org/10.1038/nj7444-259a.

Hardisty, D.J. and D.A.F. Haaga, (2008), "Diffusion of treatment research: Does open access matter?", *Journal of Clinical Psychology,* Vol. 67, No. 7.

Hey, T. and A. Trefethen (2003), "The data deluge: An e-science perspective", in F. Berman, G.C. Fox and A.J.G. Hey (eds.), *Grid Computing: Making the Global Infrastructure a Reality,* John Wiley & Sons, Ltd., Chichester, England, pp. 809-24, http://eprints.ecs.soton.ac.uk/7648/1/The_Data_Deluge.pdf.

Holocher-Ertl, T., B. Kieslinger (2013), "Towards a better society of empowered citizens and enhanced research", Green Paper on Citizen Science, Citizen Science for Europe, Socientize, Citizen Science Projects.

Houghton, J. and P. Sheehan (2009), "Estimating the potential impacts of open access to research findings", *Economic Analysis and Policy*, Vol. 29, No. 1, pp. 127-42.

Houghton, J., A. Swan and S. Brown (2011), "Access to research and technical information in Denmark", Technical Report, School of Electronics and Computer Science, University of Southampton.

Houghton, J. et al. (2010), "Economic and Social Return on Investment in Open Archiving: Publicly Funded Research Outputs", report to SPARC (Scholarly Publishing and Academic Resources Coalition), Center for Strategic Economic Studies, Victoria University.

JISC (2014), "The value and impact of data sharing and curation: A synthesis of three recent studies of UK research data centers", JISC (UK Joint Information Systems Committee), http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf, accessed 15 April 2015.

Jirotka, M. et al. (2006), "Special issue: Collaboration in e-research", *Computer Supported Cooperative Work (CSCW)*, Vol. 15, No. 4, pp. 251-55, http://dx.doi.org/10.1007/s10606-006-9028-x.

Kotarski, R. et al. (2012), "Report on best practices for citability of data and on evolving roles in scholarly communication", *Opportunities for Data Exchange*, www.stm-assoc.org/2012_07_10_STM_Research_Data_Group_Data_Citation_and_Evolving_Roles_ODE_Report.pdf, accessed 15 April 2015.

Kowalczyk, S. and K. Shankar (2010), "Data sharing in the sciences", *Annual Review of Information Science and Technology*, Vol. 45, pp. 247-94.

Lakhani, K.R. et al. (2007), "The value of openness in scientific problem solving", www.hbs.edu/faculty/Publication%20Files/07-050.pdf, accessed 15 April 2015.

Lane, J. et al. (eds.) (2014), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press.

Mooney, H. (2011), "Citing data sources in the social sciences: Do authors do it?", *Learned Publishing*, Vol. 24, No. 2, pp. 99-108.

Mooney, H. and M.P. Newton (2012), "The anatomy of data citation: Discovery, reuse and credit", *Journal of Librarianship and Scholarly Communication*, Vol. 1, No. 1, eP1035.

Murray, F.P. et al. (2009), "Of mice and academics: Examining the effect of openness on innovation", National Bureau of Economic Research, www.nber.org/papers/w14819 (accessed 15 April 2015).

Myllymäki, P. (2013), "Data to Intelligence (D2I) Research Programme on Intelligent Data-Driven Services", presentation, www.digile.fi.

Narayanan, A. and V. Shmatikov (2008), "Robust de-anonymization of large sparse datasets", SP 08 Proceedings of the 2008 IEEE Symposium on Security and Privacy, IEEE Computer Society, Washington, DC.

OECD (2015a), "Assessing government initiatives on public sector information: A review of the OECD Council Recommendation", OECD, forthcoming.

OECD (2015b), *Making Open Science a Reality*, OECD, forthcoming.

OECD (2014a), "Unleashing the power of big data for Alzheimer's disease and dementia research", Main points from the OECD Expert Consultation on "Unlocking Global Collaboration to Accelerate Innovation for Alzheimer's Disease and Dementia", DSTI/ICCP/IE(2013)13/FINAL, OECD Publishing, Paris.

OECD (2014b), *OECD Science, Technology and Industry Outlook 2014,* OECD Publishing, Paris.

OECD (2013), "New data for understanding the human condition: International perspectives", OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences, OECD Publishing, Paris.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, C(2008)36, 30 April 2008, OECD, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris.

OECD (2006), Recommendation of the Council concerning Access to Research Data from Public Funding, 14 December 2006, [C(2006)184], OECD Publishing, Paris.

Piwowar, H.A. and W.W. Chapman (2008), "Identifying data sharing in biomedical literature", AMIA Annual Symposium Proceeding Archive, pp. 596-600.

Piwowar, H.A. and T.J. Vision (2013), "Data reuse and the open data citation advantage", PeerJ 1:e175, https://dx.doi.org/10.7717/peerj.175.

Piwowar, H., R.S. Day and D.B. Fridsma (2007), "Sharing detailed research data is associated with increased citation rate", *PLOS ONE*, Vol. 2, No. 3.

Riesch, H., C. Potter and L. Davies (2013), "Combining citizen science and public engagement: The Open AirLaboratories Programme", *Journal of Science Communication,* Vol. 12, No. 03.

Royal Society (2012), "Science as an open enterprise", Royal Society Science Policy Centre Report, February.

Rowlands, I. and D. Nicholas (2005), *Institutional Repositories for the Research Sector: Feasibility Study*, National Library of New Zealand, Wellington.

Sparks, S. (2005), *JISC Disciplinary Differences Report*, Rightscom, London.

Spiegler, D.B. (2007), "The Private Sector in Meteorology – An Update", http://journals.ametsoc.org/doi/pdf/10.1175/BAMS-88-8-1272.

The Economist (2010), "Data, data everywhere", 27 February.

Uhlir, P.F. (2012), "For Attribution – Developing data attribution and citation practices and standards: Summary of an international workshop", National Academies Press, Washington, DC.

UNESCO, (2012), *Policy Guidelines for the development and promotion of Open Access*, UNESCO Publishing, Paris.

Ware, M. (2009), "Access by UK small and medium-sized enterprises to professional and academic literature", Publishing Research Consortium, April.

Ware, M. and M. Monkman (2008), "Peer review in the scholarly journals", Publishing Research Consortium, January.

Williams, H. (2010), "Intellectual Property rights and innovation: Evidence from the human genome", National Bureau of Economic Research, NBER Working Paper 16123.

Further reading

Bjork, B.-C. et al. (2013), "Anatomy of green open access", *Journal of the American Society for Information Science and Technology*, www.openaccesspublishing.org/apc8/Personal%20VersionGreenOa.pdf, accessed 15 April 2015.

Coe, D.T. and E. Helpman (1995), "International R&D Spillovers", *European Economic Review*, Vol. 39, No. 5, pp. 859-87.

Coe, D.T., E. Helpman and A.W. Hoffmaister (2009), "International R&D Spillovers and Institutions", *European Economic Review*, Vol. 53, No. 7, pp. 723-41.

Dallmeier-Tiessen, S. et al. (2011), "What scientists think about open access publishing: Highlights from the SOAP (Study of Open Access Publishing) project survey", Cornell University Library, http://arxiv.org/abs/1101.5260 (accessed 15 April 2015).

EC (2014), *Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group,* European Commission, http://lalist.inist.fr/?p=12043 (accessed 15 April 2015).

Hall, B.H., J. Mairesse and P. Mohnen (2009), "Measuring the Returns to R&D", National Bureau of Economic Research, NBER Working Paper 15622.

Ipsos MORI (2011), "Public Attitudes to Science 2011", report to the Department for Business, Innovation and Skills, May.

Jaffe, A.B. (1989), "Real effects of academic research", *American Economic Review*, Vol. 79, No. 5, pp. 957-70.

Khan, M. and K.B. Luintel (2006), "Sources of knowledge and productivity: How robust is the relationship?", OECD *Science, Technology and Industry Working Papers,* 2006/6, OECD Publishing, Paris.

Luintel, K.B. and M. Khan (2011), "Basic, applied and experimental knowledge and productivity: Further evidence", *Economics Letters*, Vol. 111, No. 1, pp. 71-74.

Parcher, J. (2012), "Benefits of open availability of Landsat data", www.oosa.unvienna.org/pdf/pres/stsc2012/2012ind-05E.pdf, accessed 15 April 2015.

van der Meulen, B. and A. Rip (2000), "Evaluation of societal quality of public sector research in the Netherlands", *Research Evaluations*, Vol. 8, No. 1, pp. 11-35.

Verspagen, B. (2004), "Innovation and economic growth", in J. Fagerberg, D.C. Mowery and R.R. Nelson (eds.), *Handbook of Innovation*, Oxford University Press, Oxford.

Chapter 8

The evolution of health care in a data-rich environment

This chapter examines how large and diverse health data sets are being used to improve population health and support patient-centred care, health system management, and human health research. Among the aspects considered are electronic health records, smart models of care, the role of social media and crowdsourcing. The chapter also looks at barriers that will need to be overcome to pave the way for widespread data-driven innovation (DDI) in the health sector, examining issues raised by the use of personal health data not discussed in previous chapters. It concludes with a list of success factors that will enable governments to provide the leadership needed to progress further toward data-driven health research and care.

> *Big data is not just a quantitative change, it is a conceptual and methodological change. It will transform the way we do science and the way we deliver care. (Rossor in OECD, 2014a)*

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

The health sector is a knowledge-intensive industry: it depends on data and analytics to improve therapies and practices. There has been tremendous growth in the range of information that is being collected, including clinical, genetic, behavioural and environmental data. Every day, health care professionals, biomedical researchers and patients produce huge amounts of data from an array of devices, including electronic health records (EHRs), genome sequencing machines, high-resolution medical imaging, ubiquitous sensing devices (i.e. available anywhere), and smartphone applications that monitor patient health. The data generated are of great value to health care and research, and it is predicted that more medical information and health and wellness data will be generated in the next few years than ever before.

At the same time, the potential to process and analyse these emerging multiple streams and large volumes of data – big data – and to link and integrate them is growing. Such data-driven innovation (DDI) can yield many benefits, including new insights into the natural history of diseases and their diagnosis, prevention and treatment, and greater opportunity for further development of personalised therapies. Indeed, there is growing evidence that big data can be leveraged to transform health care.

Box 8.1 points to four basic categories of digital data use that can bring value to citizens, care providers and the system itself. The data can help: improve patient care; manage the health system; understand and manage population and public health; and facilitate health research. However, many challenges must be overcome before the benefits from DDI in the health sector can be reaped. One of these is that EHRs are being collected in health care systems that are often fragmented, with points of care functioning as silos. Questions of privacy also have to be addressed, and skill building will be needed to analyse voluminous health data sets.

This chapter reviews the evidence for big data's potential; equally, it considers the barriers that will need to be overcome to pave the way for widespread DDI in the health sector. The first section examines factors driving greater use of large-scale health data to generate knowledge and yield new intelligence for health system management and policy making. The second section reviews national health data sources and their capacity to be brought together, so as to understand health care pathways – that is to say, patients' progress through the health care system, from their earliest to last days – and raise the quality of clinical care. The third section provides examples of how large-scale data sources are already playing a role in DDI – helping create new and "smarter" models of care, care that is more centred on the patient, and a more efficient clinical research enterprise for improved prevention and better disease management. The fourth section provides an overview of the challenges of transforming health research with big data, including the need for infrastructure and analytical tools to analyse large health data sets.

The chapter concludes with a discussion of the factors that can enable governments to provide the leadership needed to progress further toward data-driven health research and care. The chapter thereby highlights specific opportunities for fostering Alzheimer's and dementia research in response to the direct mandate from the *G8 Dementia Summit Declaration* to the OECD to "take stock of our current national incentive structure for research […] and consider what changes could be made to promote and accelerate discovery and research and its transformation into innovative and efficient care and services".[1]

Box 8.1. **Uses of digital data in the health system**

Improving patient care – Secondary use of health data can improve quality initiatives in and the effectiveness of patient care, in both clinical and home care settings. For example, administrators and front-line clinicians can be alerted when measures related to quality and patient safety fall outside a normal range, and notified of factors that may be contributing to the deviations. Clinicians can aggregate and reuse data from their patients to evaluate their own performance against clinical practice guidelines. The data can also provide insights that lead to revised care protocols.

Managing the health system – Health data can be used to manage and improve the effectiveness and efficiency of the health system by informing decisions regarding programmes, policy and funding. For example, costs can be reduced by identifying ineffective interventions, missed opportunities, and duplication of services. Access to care can be increased and wait times reduced by understanding patient journeys across the continuum of care; by ensuring that patients receive the services most appropriate for their needs; by accurately projecting the future health care needs of the population; and by optimising the allocation of resources across the system.

Understanding and managing population and public health – Health data can be used to understand the burden of illness and quality of life of the population, and to manage and evaluate public health interventions, including for health promotion and prevention. For example, in addition to timelier public health surveillance of influenza and other viral outbreaks, data can be used to identify unanticipated side effects and contraindications of new drugs.

Facilitating health research – Health data can be used to support research in many fields that informs clinical programmes, health system management, and population and public health. For example, multiple sources of data can be integrated to find early (bio)markers of disease; the comparative cost-effectiveness of different interventions can be evaluated; and historical data can be used to simulate and model trends in long-term care needs, and evaluate different policy options to meet those needs.

Source: Adapted from Canadian Institute for Health Information, 2013.

8.1. Drivers of growth of digitised health data

The amount of digitised data available in the health sector is growing rapidly. There are five principal factors driving the increased collection and use of large-scale data in this sector. They are: i) demographic changes and the shifting of the global disease burden toward long-term non-communicable diseases; ii) fiscal pressures and the need for greater efficiencies; iii) the need for more responsive, patient-centred services; iv) increasing global co-operation to address common health problems; and v) the volume, velocity and variety of health data available (see Chapter 3).

Demographic changes and non-communicable diseases

The global burden of disease has shifted in the past 20 years, from infectious conditions to long-term non-communicable diseases (NCDs) brought on by lifestyle choices and environments: heart disease, stroke, diabetes, chronic neck and back pain, cancer, and depression (IHME, 2013). Further, populations are ageing and many people are living longer with multiple morbidities (concurrent diseases) and disabling conditions. Since 1970, average life expectancy has risen by 35 years worldwide, with gains in years achieved across the world's regions. OECD countries have witnessed extraordinary gains

in longevity, with average life expectancy at birth rising 10 years since 1970 to exceed 80 years in 2011 (OECD, 2013a). Globally, the proportion of people over 80 years of age in particular is anticipated to increase by 2.5 times between 2010 and 2050 (UN, 2013). The rapidly expanding cohorts of elderly and older elderly will include a significant proportion of persons with chronic diseases.

The rising burden of NCDs, multimorbidities and risk factors for NCDs has important implications, for how care is best organised and provided; where new treatment innovations and preventive approaches can be expected; and future cost pressures.

To address that burden, medicine must focus on preventing the onset of NCDs and controlling their progression, including through lifestyle changes. At the same time, health systems must focus on improvements in care co-ordination and delivery. The current pace of innovation in genomics, biological systems, and information and communication technologies (ICTs) has the potential to increase our ability to predict and prevent disease and promote healthy behaviours; develop cost-effective therapies; redesign health care systems to assure integrated and co-ordinated care; improve safety and quality for patients; and extend healthy lives. As will be explored further in this chapter, these advancements are closely linked to the generation and analysis of data that permit the study of full populations and their health care experiences and outcomes.

Fiscal pressures and the need for greater efficiencies

Fiscal pressures will continue to push governments to seek greater efficiency, accountability and quality in the health care sector. During the fifty years prior to 2009, health spending in OECD countries outpaced economic growth, resulting in an increasing share of GDP allocated to health. By 2009, 9.6% of GDP in OECD countries was allocated to health, up from under 4% in 1960 (OECD, 2013a). Average annual growth in health spending in real terms between 2000 and 2009 was 4.1%, compared to GDP growth of only 1.5% (OECD, 2013b). Since 2009, many countries have reduced budgets for health in response to the economic downturn. By 2012, health expenditures accounted for 9.3% of GDP (OECD, 2014b). Governments under pressure to protect funding for acute care have been cutting other expenditures, such as public health and prevention programmes. In 2012, on average across OECD countries, only 3% of health budgets were allocated to prevention and public health programmes in areas such as immunisation, smoking, alcohol, nutrition and physical activity. The long-term wisdom and sustainability of such budget reductions in spending on prevention is uncertain, as is the ability of governments to continue to contain rising costs.

Continuing pressure to find ways to make systems more productive has moved the focus from cost containment to performance-based governance. To evaluate health sector performance, managers and governments will need timely and accurate information about the prices and volumes of services provided and the health outcomes produced, at levels sufficiently detailed to take corrective policy action. The need to manage health system outcomes more actively will lead to greater use of clinical and administrative data to assess the comparative effectiveness of therapies and services. These data will also be needed for redesigning and evaluating new models of health care service delivery.

The need for more responsive, patient-centric services

The role of patients in the care process – managing their own health – has taken on much greater importance in recent years. In order to address patients' expectations for seamless care, it will be increasingly important to improve co-ordination and integration

of care provided by different parts of the health and social care systems. Patients' taking command of the management of their own health will support better outcomes and coordination of care, particularly for patients with chronic diseases who often require services from multiple health care providers.

The increasing use of electronic medical records promotes patients' participation in their care, self-management of health conditions, and informed decision making. Patients' interest in their diagnostic test results and medical records, in their options for care, in the quality of providers, and in scheduling visits online will keep growing. Over the past decade multiple studies have documented the value of electronic personal records (EPRs) in supporting greater patient-centred services. As an illustration, the US Department of Veterans Affairs offers patients access to an EPR that includes details from their medical history, such as clinical notes and laboratory test results. A patient survey to evaluate the service indicated that the majority of veterans viewed the personal record as helpful to them; as having made it easier to locate information they needed; as having improved their care; and as a tool they would recommend to others (Nazi et al., 2013).

Patients and practitioners are also increasingly interested in devices, tools and computer applications that assist in monitoring and improving health and well-being. They recognise that these can help patients live longer in their own homes rather than in considerably more expensive hospital or nursing home facilities; enable longer-term independent living; and encourage personal responsibility for healthier lifestyles (OECD, 2013c). Many such emerging information and communication systems have the potential to provide new streams of data for evaluating treatments and measuring and evaluating health care outcomes. However, in many countries challenges have yet to be met to unleash this potential.

Increasing global co-operation to address common health problems

The fourth driver is the need for co-operation to tackle global public health challenges such as infectious diseases, and improve early detection and warning of emerging health threats and events. Examples include the Program for Monitoring Emerging Diseases (ProMED), established in 1993, which has demonstrated the power of networks and the feasibility of designing effective, low-cost global reporting systems. ProMED has also encouraged the development of additional electronic surveillance data-sharing networks – such as the Global Public Health Information Network (GPHIN)[2] and HealthMap.[3]

Influenza surveillance is one of the most developed global surveillance and monitoring systems of the World Health Organization. It began in 1948 and has developed over the years into a highly successful global partnership. The network now involves 110 collaborating laboratories in 82 countries, constantly monitoring locally isolated influenza viruses and providing real-time streams of data on the emergence and spread of different strains.

Complementing these traditional case-based and syndromic surveillance systems, monitoring of unstructured events – through news and Internet media, web searches (e.g. Google Flu trends), etc. – has been a significant component of public health early warning and response over the past decade. More recently, with the increase of Web 2.0 platforms and social media, there is a new real-time source of intelligence provided by citizens that is immediately in the public domain and thus readily available. During the recent Ebola outbreak in West Africa, epidemiologists and telecommunication companies were exploring the potential of new data sources, such as mobile phones, to better model the spread of the disease (The Economist, 2014; Wall Street Journal, 2014).

In addition to monitoring, there is increasingly global interest in research to tackle the emergence of NCDs, through better preventive interventions and treatments. The OECD is actively engaged in a global project to improve data sharing and access internationally to accelerate innovation that addresses dementia (OECD, 2013d). The focus on dementia is the result of a direct mandate from the *G8 Dementia Summit Declaration* to the OECD.

Volume, velocity and variety of health data

The fifth and possibly most important driver of health data use is the sheer volume, velocity and variety of health data available. As will be discussed in the next sections of this chapter, many health care systems are rapidly digitising immense amounts of clinical, financial and operational data and using them for a wide range of activities, including:

- preventive care, e.g. early detection
- field data to support emergency and urgent care
- coaching, rehabilitation and maintenance
- intervention, e.g. reminders
- epidemiological assessments
- post-market surveillance and analysis
- health care quality and performance monitoring.

This will require real-time continuous archiving of multi-modal data sets and multi-domain collaborative annotations, as well as post-therapeutic visualisation of the archived data. The volume of this data is set to increase dramatically with advances in mHealth (mobile health, involving mobile devices), sensor and imaging technologies to support diagnosis and treatment. Further, these data are heterogeneous (structured, unstructured, text, etc.), reflecting the traditional silos across care settings, industry/research, and scientists/clinicians. The cost and complexity of linking data stored in these various formats are decreasing, enabling analysis of health care interventions and utilisation enhanced with additional information about personal behaviours, lifestyles and genetic profiles.

This remarkable expansion of digital health data is in turn largely driven by the confluence of important technological developments. These include notably the increasing ubiquity of broadband access and the proliferation of smart mobile devices and emerging smart ICT applications, empowered by sensor networks and machine-to-machine (M2M) communication. Cloud computing has also greatly increased data storage and processing capacity (see Chapter 3). Great reductions in storage costs over the past 20 years have also been a significant driver, as they have enabled the collection and use of large volumes of health-relevant data; electronic health records and genetic, neuroimaging and epidemiological data are just a few examples.

All of these drivers have greatly increased not only the availability of data in the health sector, but also – with developments in computing power – their use, creating new opportunities to obtain insights. The rise of chronic conditions and fiscal pressures will make it increasingly important to be able to follow health care pathways and determine which paths deliver better outcomes in an efficient manner. Patients will want the health sector to improve therapies, and will want health care experiences to be as modern as other business services in terms of service responsiveness, transparency and

communication. Opportunities for global co-operation in sharing data to find solutions to common challenges may continue to present themselves, particularly as the urgency of addressing NCDs and new infectious diseases rises. For all of these reasons, there will be continued interest in developing and using data to advance health care therapies, health care delivery, and health system governance. The next sections review the four basic categories of data use described in the introduction and then discuss the critical success factors and policy priorities for addressing challenges that may be limiting data sharing and use.

8.2. Data-driven innovation to improve health care quality and health system performance

Essential to the monitoring and evaluation of both health care quality and health system performance is the ability to track patients as they progress through the system, from primary health care to speciality care to hospitalisations, long-term care, home care, hospice care and death. Tracking data should also provide information about patient characteristics, illnesses, medications, therapies, laboratory tests and medical images. This type of follow-up permits a comprehensive view of health care services and evidence of what is effective and under what circumstances. It can also uncover, among other things, medical errors, adverse drug reactions, fraud, adherence to clinical guidelines or lack thereof, optimal care paths and patients with optimal treatment results.

Although the capacity to collect and analyse data related to health care pathways is increasing, only a handful of countries have health information systems organised to permit comprehensive views of patient care across the health care continuum. Key pieces of information about patients' care paths are instead often isolated in various databases, such as hospital admissions and discharges, primary care records, insurance claims, pharmaceutical databases, image banks and patient surveys.

This section begins with a review of current national health data assets and their capacity to be integrated to understand health care pathways. It goes on to discuss the development and use of electronic health record systems and their potential to improve measurement of health care pathways. Examples are provided of how some countries have developed data to follow the pathway of care in order to monitor health care quality and health system performance nationally, and to contribute to international comparisons. It then describes the emerging field of comparative effectiveness (or relative effectiveness) research, which uses health care pathway data to determine which therapies or processes of care are the most effective.

Key national health data sets

While countries are investing in data infrastructure, a 2011/12 OECD survey of national health data sets (OECD, 2013e) reveals significant cross-country differences in data availability and use. Some countries have seen significant progress in data use and its compatibility with robust privacy protection. Others have limited data and restrictions preventing access to its use, even by government (see Table 8.1). Most of the countries responding to the survey reported that they collect national data across the continuum of health care services and at the level of individual patients or persons. Users of national data included governments, insurers, research institutes and health care providers. About half of the countries reported that some of their key data sets were regularly linked together for research purposes or statistics to better understand health care pathways and outcomes. Similarly, about half of the countries reported that some of their key data sets

were routinely linked to monitor health care quality. Very few countries, however, link data routinely to monitor health care quality in several important areas of health care: prescription medicines (seven countries); mental hospital in-patients (five countries); primary health care (four countries); and long-term care (four countries).

Table 8.1. **Number of countries[1] reporting data and data linkages**

Data is about:	Hospital in-patients	Deaths	Cancers	Rx[2]	Mental hospital in-patients	Primary care	Long-term care	Health risks and behaviours	Socio-economics: income, education, employment, ethnicity
National data set available…	19	19	17	14	17	16	16	19	19
Contains records for patients or persons	16	17	16	12	14	13	13	16	16
Is linked to other data for health research or monitoring	14	15	13	12	8	10	11	10	11
A linkage study is usually[3] under way	12	15	11	10	7	8	6	7	11
A linkage study to monitor *health care quality* is usually under way	12	12	11	7	5	4	4	4	4

1. Nineteen countries responded to the survey. Australia, Belgium, Canada, Denmark, Finland, France, Germany, Israel, Japan, Korea, Malta, Norway, Poland, Portugal, Singapore, Sweden, Switzerland, the United Kingdom and the United States.

2. Pharmaceutical drug utilisation.

3. The data set is used to undertake record linkage projects on a regular basis, such that a data linkage project involving the data set is usually underway.

Source: OECD, 2013e.

Electronic health record systems

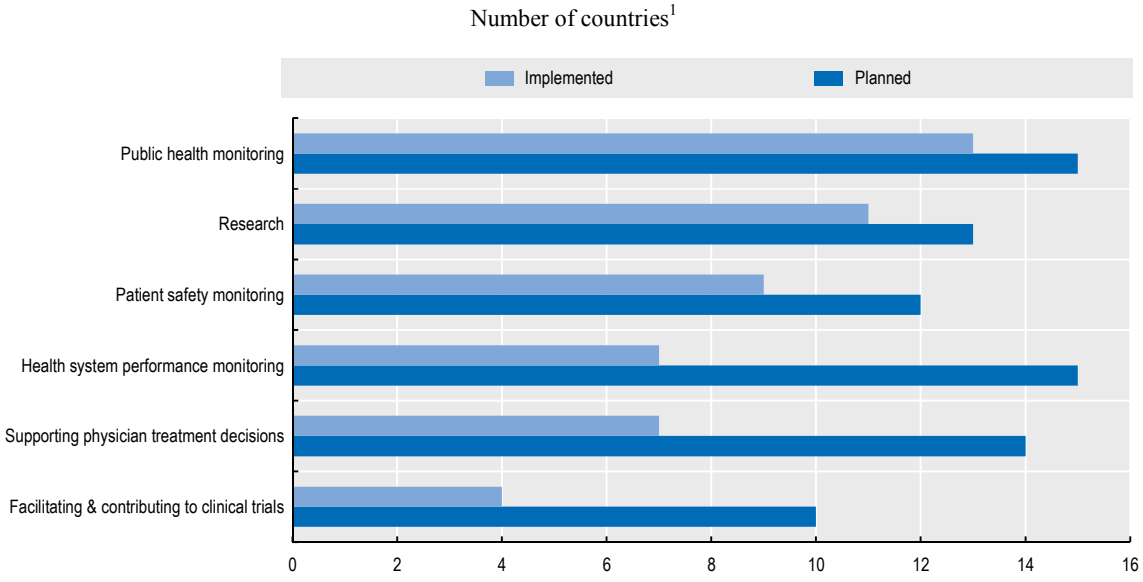
The development and use of data from electronic health records (EHRs) have the potential to support health care DDI and to improve the quality, safety and performance of health care systems. In its most mature form, an EHR system contains or virtually links together information about a patient's health care encounters, including diagnosis, radiology, laboratory tests and medications. Patient identifying information is necessary to bring the information together from various related data sets and then to retrieve the information when it is needed. As time passes, a comprehensive health care biography can emerge from the data available in the system to support the care of the individual patient; for population-level statistics and research, it can improve existing therapies, discover new ones, and improve the quality, safety and performance of health care systems.

In 2012 most of the countries studied (22 of 25) reported a national plan or policy to implement EHRs, and most had already begun to implement that plan (20 countries). EHR systems in some countries include data on patients' key characteristics and health problems, as well as their history of encounters with the health care system and the treatments they have received from a variety of health care providers. The greatest

contribution of these systems, as they develop, is the potential for secondary analysis of the data to monitor and conduct research to improve the health of the population and the quality, safety and efficiency of health care.

Of the 25 countries studied, 18 had included some form of secondary analysis of EHRs within their national plan (Figure 8.1). The most commonly included secondary uses reported were public health and health system performance monitoring. Fourteen countries also indicated that they intended for physicians to be able to query the data to support treatment decisions. The least commonly reported planned data use was for facilitating or contributing to clinical trials. This use was noted by ten countries.

Figure 8.1. **Planned and implemented uses of data from electronic health record systems**



1. Twenty-five countries responded to the survey. Austria, Belgium, Canada, Denmark, Estonia, Finland, France, Germany, Iceland, Indonesia, Israel, Japan, Korea, Mexico, Netherlands, Poland, Portugal, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom and United States.

Source: OECD, 2013e.

Many countries also reported that regular use of EHR data for public health monitoring (13 countries) and general research (11 countries) was already under way. There are currently several ongoing projects addressing the (re)use of EHR data for purposes of clinical research. In the United States, initiatives such as i2b2,[4] the eMERGE network,[5] the Kaiser Permanente Research Program on Genes, Environment and Health (RPGEH)[6] and the Million Veteran Program[7] are focusing on integrating EHRs and genomic data (Jensen, Jensen and Brunak, 2012). The Stanford Translational Research Integrated Database Environment (STRIDE) is an example of a US project that aims to create an informatics platform supporting clinical and translational research.[8]

In Europe, a number of research projects and initiatives such as the i4health network,[9] EMIF (European Medical Information Framework),[10] eTRIKS (Delivering European Translational Information & Knowledge Management Services),[11] INTEGRATE (Integrative Cancer Research through Innovative Biomedical Infrastructures),[12] Linked2Safety,[13] SALUS (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies)[14] and TRANSFoRm (Translational

Research and Patient Safety in Europe)[15] are concerned with (re)using EHRs to facilitate clinical research by focusing on different disease domains and addressing different data use cases and scenarios.

There are several significant differences between countries whose national plans or policies called for at least four of the data uses outlined in Figure 8.1 (the engaged) and those who were planning on fewer or no secondary data uses (the cautious). Data privacy and security concerns have been one of the major barriers to the adoption of EHRs and implementation of a national health information exchange in a number of countries. These issues will be reviewed in some depth in the last section of this chapter. Other barriers include a lack of technology standards that could facilitate interoperability, and the cost of implementing such a system. Engaged countries were somewhat more likely than cautious countries to report having created national governing bodies responsible for clinical terminology and interoperability standards – 62% compared with 50%. Terminology standards ensure that the data is captured in a consistent manner through a structure that enables statistics and analysis. Interoperability standards ensure that records can be shared or exchanged. Nonetheless, where health care systems are fragmented, with points of care operating in silos, implementation of standardised EHR systems remains challenging.

Virtually all of the engaged countries (92%) have developed a national minimum data set that standardises the content of patient records that are intended to be shared among health care providers. In contrast, only one-half of the cautious countries have defined a minimum data set. Engaged countries (54%) are also somewhat more likely than cautious countries (42%) to report that their EHR system is already being used to create data sets for statistics and research. As a result, they are also more likely to have instituted processes for auditing the clinical content of electronic records for quality, although this is still relatively rare for both groups. Many engaged countries have also consulted with the public on privacy and security issues, and have developed data governance frameworks that permit privacy-protective data uses.

Harnessing value from data to improve health system performance

Countries that are actively monitoring health care quality and health system performance provide very interesting examples of how the data are being used and the benefits accrued. Examples of data use range from evaluation of the quality and cost-effectiveness of treatments to monitoring adverse events related to pharmaceuticals and medical devices; incorporating the results of care pathway analysis into evaluations of and revisions to clinical care guidelines; and building pathway data to promote world-class research.

Evaluation of treatment quality and cost-effectiveness

Finland monitors the content, quality and cost-effectiveness of a set of selected diseases and treatments (stroke, premature newborns, hip fracture, breast cancer, schizophrenia, heart attack, hip and knee replacement surgery, and invasive heart surgery) by linking patient data for the Finnish population across the whole cycle of care, from admission to hospital to care by their community doctor to the medications prescribed and deaths (OECD, 2013e). From both administrative data and data extracted from electronic health records, Finland has new indicators for each hospital to evaluate treatment quality and cost, including: mortality rates; emergency room visits and readmissions to hospital;

infections and complications; and stays in nursing homes and home care visits. Hospital quality is improving, as the results are publicly available.

Within the *United Kingdom, England* has a new initiative called care.data that aims to create data about episodes of care. Included are both health care and social care, with data pertaining to pathways between primary and secondary care and information about diagnosis, laboratory tests and prescription medications (NHS, 2013). The six aims of the care.data initiative are to support patient choice, advance customer services, promote greater transparency, improve outcomes, increase accountability, and drive economic growth by making England a centre for world-class health services research. Data for consenting patients within the entire population of England will be linked, with data extracts taking place monthly to ensure timely monitoring.

Japan has created a new medical insurance claims database to assist the Ministry of Health, Labour and Welfare in the preparation, implementation and evaluation of a plan to optimise medical care costs. The data were provided to researchers and to prefectures on a trial basis in 2011 and 2012. Several cost and quality studies were undertaken and published as a special issue of the *Journal of the National Institute of Public Health*. These studies included a linkage of insurance claim data with data on the provision of guidance to patients during periodic health check-ups regarding metabolic disease (Okamoto et al., 2013). The study found a reduction in the onset of metabolic disease and in health care expenditures among patients who received guidance about reducing disease risk during health check-ups.

Monitoring the underuse, overuse and misuse of therapies

Korea uses population-wide health insurance claim data to identify underuse, overuse and misuse of therapies and to reduce variation in care practices by regularly reporting quality indicators, including mortality and readmission after hospital procedures; inappropriate prescribing in primary care; and outcomes following discharge from mental health hospitals (OECD, 2013e). Korea links claims data for patients across the entire pathway of care, and is able to report timely results.

Quality and efficiency assessments of clinical care guidelines

Sweden is breaking new ground by using data to undertake both quality and efficiency assessments of clinical care guidelines (OECD, 2013e). These guidelines inform physicians and health care professionals about the most appropriate therapies for patients with different health profiles and problems. By following a patient's cycle of care, they are able to evaluate the extent to which guidelines are being followed and whether or not the health outcomes of the patient meet expectations. This evidence is then used to revise the guidelines, completing an ongoing cycle of improvement in care quality and efficiency.

Monitoring adverse events related to pharmaceuticals and medical devices

The *United States* Food and Drug Administration has implemented a sentinel project to transform how it monitors the safety of the medicines, medical devices and biologics that it regulates, by tapping directly into electronic health records, administrative data and insurance claim records. Building toward a nationwide rapid-response electronic safety surveillance system, the sentinel pilot study involves 17 data partners across the United States, and encompasses the data of nearly 100 million patients (FDA, 2013).

The *EU* Advanced Drug Reporting (EU-ADR) [nowhere do I find that name linked to the abbreviation. ADR instead would appear to stand for "Adverse Drug Reactions" – please confirm] initiative defined a proactive strategy for post-market drug assessment based on automating analysis of data stored in large electronic health record databases in four European countries (Denmark, Italy, the Netherlands and the United Kingdom) and covering 30 million patients (Coloma et al., 2012). EHR data are analysed to identify a ranked list of signals of potential adverse events and their significance in terms of health risks. Adverse events monitored include acute myocardial infarction, acute renal failure, anaphylactic shock and gastrointestinal bleeding. Results indicate that active surveillance for signal detection with health care database networks is feasible, but that it would be necessary to expand the data network coverage to a larger pool of patients – that is, to more participating countries – to monitor the effects of infrequently used drugs.

Generating clinical pathway data to promote health services research

In the *United States*, Kaiser Permanente, a health care maintenance organisation (HMO) with 8 sites and 9 million members, has 7 research centres conducting public domain research with patient-level data. Kaiser's experience with linking data across the health care pathway for research extends back 50 years. Kaiser is now at the forefront of this field with the data it can extract from its electronic medical system and the data it can link together with patient care pathways from its biobank. A new study that Kaiser described to the OECD involves examining whether certain prescription medicines for mental illness may be linked to the development of genetic mutations in humans (OECD, 2013e).

In *Canada*, the Institute for Clinical and Evaluative Sciences (ICES) is a research centre at the University of Toronto that provides population-based health services research for Canada's largest province, Ontario (OECD, 2013e). ICES collects personal health data from the Ontario Ministry of Health and Long-term Care and other entities. Findings in 2013 included that commonly co-prescribed statins and antibiotics are linked to muscle loss and kidney failure in seniors; that a recent colorectal cancer screening programme was not able to fully address inequities in access; and that implementation of North America's first stroke care facilities improved outcomes (ICES, 2013).

Cross-country comparisons of health system performance

Collaborative big data efforts to improve health system performance investments in the development of internationally comparable population-level health data are leading to new ways to benchmark and compare how health systems are performing to help countries to improve patient safety, health outcomes and system performance. Within Europe such efforts are funded by the European Union. Two examples from the EU Seventh Framework Programme are EuroHOPE and ECHO.

EuroHOPE – the European Health Care Outcomes, Performance and Efficiency Project – is evaluating the performance of European health care systems in seven countries, in terms of outcomes, quality, use of resources and costs (Häkkinen et al., 2013). Participating countries include Finland, Italy, Netherlands, Norway, Sweden and the United Kingdom (Scotland). Health care data for hospitalisations, pharmaceuticals, registered cancers and deaths are linked to follow patient pathways of care. The patient groups studied are those with acute myocardial infarction, stroke, hip fracture, breast cancer and low birth weight. EuroHOPE is developing indicators that it will recommend to the European Union for routine reporting; developing methods for international comparative health services research based on the linkage of person-level data; and

informing the public about the policy-relevant drivers of health care quality – including treatment practices, use of medicines and new medical technologies, waiting times, organisation of care, and costs.

ECHO, the European Collaboration for Healthcare Optimization project, has pooled hospital administrative and contextual data from seven countries (Austria, Denmark, England, Portugal, Slovenia, Spain and Sweden) to learn more about variation in care access and outcomes, and the relationship between this variation and the socio-economic status of the areas in which patients live.[16] ECHO intends to explore whether place of residence and access to particular health care providers have a bearing on whether or not care is safe and effective, by examining within-country and between-country variations. ECHO is the first international health system performance comparison to pool personal health data, into a data set of 200 million hospital discharges.

Assessing variability in health care treatment across countries

In terms of single disease areas, there is no doubt that the long history of databases to register cases of cancer has endowed research in that area with the evidence necessary to monitor and advance quality of care. Indeed, of the 19 countries responding to the OECD survey in 2011/12, 16 were maintaining national cancer registry data sets. There is also a long practice in most OECD countries of linking cancer registrations and death databases in order to estimate cancer survival rates. The CONCORD-2 study is a worldwide comparison of cancer survival from over 270 cancer registries in 61 countries for 10 cancer sites in adults and childhood leukaemia. This study examines underlying causes of differences in survival rates.[17] The International Cancer Benchmarking Project is advancing this research further. In this project, cancer registries with details about the cancer stage at diagnosis have been analysed in six countries (Australia, Canada, Denmark, Norway, Sweden and the United Kingdom), to compare differences in survival and to discover why differences occur. Thus far, the researchers have found that patients in Sweden are the most likely to survive at least one year after diagnosis of breast, bowel and lung cancers; those in the United Kingdom are the least likely (Cancer Research UK, 2013). The role of treatment in survival differences by cancer stage is the next stage of inquiry for the project.

The emerging field of comparative effectiveness research (CER)

Comparative effectiveness research is designed to inform health care decisions by providing evidence on the effectiveness, benefits and harms of different treatment options. The evidence is generated from research studies that compare drugs, medical devices, tests, surgeries, and ways to deliver health care (AHRQ, 2013). CER ultimately seeks to provide pragmatic knowledge that can be applied toward delivering "the right treatment to the right patient at the right time". Achieving this goal in an area as complex as health care, however, requires robust, accessible data sources capable of providing detailed patient-level information in a time- and cost-efficient fashion.

Patient registries have been used throughout Europe for CER as well as for patient-centred health outcomes research (PCOR). A recent review of disease and treatment registries in Europe (Larsson et al., 2012) describes how improvements in health outcomes, like fewer revisions after hip replacement surgery and gains in survival after acute myocardial infarction (AMI), followed implementation of public reporting of outcomes by providers and the engagement of the clinical community to address quality concerns.

Two incentive programmes in the United States have increased the volume and velocity of data generated for CER in recent years (*The Daily Briefing*, 2011). The first was an allocation of USD 1.1 billion for CER funding within the 2009 federal stimulus package. The second was the launch of the Patient-Centered Outcomes Research Institute in 2011, with the capacity to fund USD 550 million to conduct CER and establish priorities for national CER (PCORI, 2013).

Three representative examples of the CER approach have been published in the United States. The first is an evaluation of a comprehensive programme to control hypertension; the second is an evaluation of methods to improve colorectal cancer screening. The third identified an unexpected and dangerous interaction between two of the most widely prescribed medications: pravastatin, prescribed for hypertension control, and paroxetine, an anti-depressant (Jaffe et al., 2013; Mosen et al., 2010, 2013; Tatonetti et al., 2011).

The rollout of EHRs in many OECD countries will help health care systems reach the vision of CER as a valuable resource for informed health care decision making.

8.3. Data-driven innovation for smarter models of care

There is a broad and growing consensus that any systematic effort to address today's health and wellness challenges will also require data to support new and "smarter" models of care, that recognise the need to keep the elderly and the disabled in their own homes rather than in the considerably more expensive hospital or nursing home systems; that enable longer-term independent living; and that encourage personal responsibility for healthier lifestyle choices. The effort will require enhanced capacity for the sharing, processing and analysis of health and behavioural data to support patient-centred care, and a more efficient clinical research enterprise for improved prevention and better disease management.

Taking shape alongside these goals is a vision for a "learning health system". The Institute of Medicine, a long-time proponent of the concept, defines a learning health system as: "... one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care" (Grossmann, Powers and McGinnis, 2011).

One example of a rapid learning health care system is the American Society of Clinical Oncology's Cancer Learning Intelligence Network for Quality (CancerLinQ) system,[18] CancerLinQ is designed to address the growing challenge of managing the deluge of data emerging from precision medicine for cancer care. The system incorporates data from researchers, providers and patients in order to continually improve comprehensive clinical algorithms reflecting preferred care at a series of decision nodes for clinical decision support.

These concepts, and the new models of care they represent, require a major shift from traditional practices. Today's care is reactive, episodic and focused on disease. The new health care will need to be proactive, preventive and focused on quality of life and well-being.

Current health care is usually provided within hospitals and clinics. New smart models of care could become more patient-centric, with greater opportunity for care to be

provided at home and include the broader social network (family and community) as a significant contributor to individual health and well-being.

Smarter models will be data-driven and promise to deliver greater safety and efficacy through evidence-based approaches and personalised care. This section examines the strategic directions that OECD countries are considering to realise this vision for health and wellness, from both the technological and policy viewpoints. It looks at the role of big data and ICTs, and discusses the research and policy options that could further the development of smarter models of care. It draws on the OECD – United States National Science Foundation workshop (and related report) entitled "Building a Smarter Health and Wellness Future", which was held in Washington, DC on 15-16 February 2011.

Personalised care

The power of health information processing is such that it is possible today to personalise therapy in wholly new ways. Culture, living style, belief systems and expressed choice comprise one dimension. A second dimension is the ability to search and process electronically recorded medical histories of individuals. This enables rapid identification of not just personal biological responses such as allergies, but also a much richer pattern of personal information such as results of diagnostic tests and outcomes of particular therapies. Thirdly, new genomic knowledge can help identify population group variations that influence care response, but also personal genetic profiles that can inform not just individual therapies but also selective targeted prevention.

Advances in DNA sequencing and whole genome analysis have made it possible to develop a greater understanding of response to treatment. In oncology, for example, pathologists measure whether the cancer is hormone sensitive to determine eligibility for tamoxifen therapy among those suffering from breast cancer. Effectiveness has been found to be contingent on an enzyme (cytochrome enzyme P450 2D6) needed to metabolise the drug, although the results have not always been consistent across studies (Roederer, 2009).

With the costs of whole genome sequencing declining, the expectation is for personalised medicine to be streamlined into medical practice. New data management and processing methods are needed in four areas to realise this potential: i) the processing of large-scale robust genomic data; ii) interpreting the functional effect and the impact of genomic variation; iii) integrating systems data to relate complex genetic interactions with phenotypes; and iv) making the data available at the point of care in such ways that the comprehensiveness of the information provided to the clinician supports the clinician's ability to accurately and rapidly prescribe drugs that are safe and effective for a specific patient (Fernald et al., 2011).

In addition, the full extent of patient data will need to be accessible so that questions spanning multiple data sources can be asked and answered. The consistency and completeness of patient EHRs will be increasingly important.

Ubiquitous and pervasive patient care

The ubiquitous care model is based on the utilisation of smart sensing and biometric devices for real-time monitoring, analysis and transmission of health data. The information can then be accessed by health care providers for informed diagnosis, clinical decisions regarding treatments, and evaluation of outcomes. It can also be viewed and acted upon by patients for both education and prevention.

The technology to support ubiquitous sensing already exists, and today an increasing amount of physiological monitoring data streams are displayed on medical devices. The key challenge is to combine these technologies with network infrastructure to create an integrated architecture that extends care outside the hospital to the home and to mobile patients – thus the term ubiquitous.

For example, in the case of managing patients with acute diabetes, the blood glucose level can be monitored continuously through an implant that controls the insulin delivery from a reservoir. In cardiology, there is increasing recognition of the value of implantable sensors for continuous monitoring of the most important physiological parameters for identifying the precursors of major adverse cardiac events, including sudden death. The data streams provide enormous potential for improved diagnostics, prevention, support of evidence-based practices, and remote health care. These data can yield answers to clinical questions, or raise new questions that influence care responses.

An area that is progressing rapidly is the Body Area Networks (BAN). Medical applications of BAN cover continuous real-time sampling of biomedical signals, monitoring of a person's vital signal information, and of low power medical devices. They can be broadly classified into two categories depending on their operating environments. One is the so-called wearable BAN, which is mainly operated on the surface or in the vicinity of body, such as medical monitoring.

Another is the so-called implantable BAN, which is operated inside the human body, e.g. the capsule endoscope and pacemaker. The former provide long-term health monitoring of patients in natural physiological states without constraining their normal activities. The latter allow communication between implanted devices and remote monitoring. One example of smart application is the "virtual ward", in which patients are monitored at home and visited by mobile medical teams when the data show that it is necessary. That is generally better for the patients and may be less expensive.

Ubiquitous computing can also be leveraged as a means to provide context and location-aware cues for health action. The greatest power of such techniques comes from the capacity to cross-link information drawn from multiple sensor systems and other information sources. For example, GPS data can cross-link with accelerometer-based physical activity estimates and geographic information systems (GIS). Local area communication standards, such as Bluetooth, can be used to determine the relative proximity of individuals to each other or fixed locations, relevant to the study of infectious disease.

Cross-linked sensor-based information can be used to persuade individuals to perform health behaviours – examples include encouraging people to take the stairs instead of the elevator in order to increase physical activity levels, and using text messages on a mobile phone to remind a person to measure their blood glucose. Ubiquitous sensors therefore have a particularly strong role to play in integrating health care by providing clinicians a novel and less-biased window into the habits and behaviours of their patients. This of course comes at a cost to individual privacy, and decisions must be based on voluntary and freely given consent.

The ever decreasing cost of sensor-based smart devices, together with the medical need for better information regarding a patient's habits outside the clinical environment, makes widespread adoption of these systems not only possible but indeed probable. Properly validated, these sensors have the potential to transform both personal and institutional care by providing reliable contextual information to individuals and practitioners.

Mobile health for greater patient engagement

Smarter health and wellness must address not only change in health care delivery, but also ways of engaging and informing the patient so they effectively achieve better health outcomes.

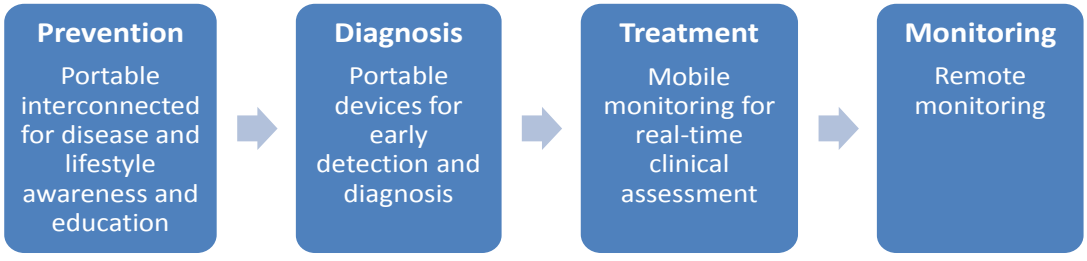
Advocates of patient-centred health have long argued for the citizen taking responsibility for their own health. This argument today applies to the prevention and management of chronic diseases such as diabetes and obesity, and health systems increasingly see their roles as agents of support. The chances of success for any prevention or care programme will depend on patient engagement and meaningful co-ownership and co-production of healthy behaviours. Indeed, a growing body of literature shows that when patients are engaged in their health care, that commitment can lead to measurable improvements in safety and quality (Dentzer, 2013; Laurance et al., 2014).

mHealth technologies can, for example, help to "nudge" people toward better decision making and remaining engaged in their care, although there are equity and safety issues that health professionals must bear in mind when recommending the use of these technologies.

Nonetheless, by putting the patient at the centre of health care transactions, health care providers can begin to overcome the silos within both specialty-based medical care and the various disciplines involved in alternative care. This requires a patient-centred data system, where every patient is a data point from which much can be learned.

mHealth offers a wide range of smart modalities by which patients can interact with health professionals, or with systems that can provide helpful real-time feedback along the care continuum, from prevention to diagnosis, treatment and monitoring (Figure 8.2). mHealth is of particular value in managing health conditions where continuous interaction is important, such as diabetes and cardiac disease. A wide range of devices is utilised for mHealth, including inter alia mobile phones, tablets, global positioning system (GPS) devices, mobile tele-care devices and mobile patient monitoring devices.

Figure 8.2. **Smart mHealth applications**



Source: OECD adapted from PricewaterhouseCoopers, 2012.

Among these devices, mobile phones in particular offer the potential to broadly and cheaply diffuse more intensive self-monitoring, feedback, self-management and clinical support than has been possible previously. This is especially true of smartphones, which support a diverse set of data streams and monitoring activities: automated traces of body movement, location, and other data that can infer physical activities, sleep, and the environment; automated and manually entered physiological measures (e.g. readings from a glucose meter); and prompted and user-initiated self-reports of the user's symptoms or behaviours. This information, appropriately managed, can be leveraged to trigger highly

personalised interventions, and thus significantly improve an individual's ability to understand and manage his or her own behaviours.

Five issues are, however, key to the successful widespread adoption of mHealth: i) establishing and sustaining engagement among participants; ii) increasing ICT knowledge across the society; iii) wide acceptance of privacy and security standards for personal data collection, analysis and use; iv) integration and interoperability – the new range of mobile devices have to function seamlessly and adapt to multiple user needs in the health sector; v) financing and new business models: there is a need to adapt regulatory structures and align incentives at different levels of the health delivery system to encourage investment in, and use of, mHealth.[19]

To achieve widespread use, mobile and health care industries will need to work toward interoperable solutions that enable economies of scale. Without agreed standards and connectivity for information exchange across the ecosystem of personal mobile devices and care services, there will be wide variation in the granularity and quality of the information collected and analysed and limited clinical utility, and payers will be reticent to invest. It is important to note, however, that individuals' engagement in mobile health requires a certain level of literacy and digital skills. Those with fewer of these skills, and who already experience poorer health conditions than those with higher levels of skills, could be excluded from using mobile health technologies, or could be at risk for not using them properly. Therefore, it is crucial for health professionals to ensure that people who are using these services either have the skills needed, or have access to opportunities for improving their skills to use the technologies effectively. Last but not least, with regard to business models, issues of cost to users of the applications as well as Internet access/bandwidth should also be considered. If not addressed properly, mobile health could create additional disadvantages for people of lower socio-economic status and/or those who live in regions with limited Internet access.

8.4. Transforming health research with big data

Big data is an integral part of the health research landscape, and indeed has even helped shape it (see Box 8.2). The next sections list a range of examples of how big data analytics offers new and more powerful opportunities to measure various aspects of disease progression and health for improved diagnosis and care delivery, as well as translational and clinical research.

Box 8.2. **Advances in genetic sequencing**

A remarkable example of the effect of big data is how, over the past two decades, the power of genetic sequencing has increased by one million-fold. No previous technology in history has increased in power that fast. DNA sequencing machines can now read about 26 billion characters of the human genetic code in less than a minute, and the sequencing cost per genome has dropped by 60% a year on average from USD 100 million in 2001 to less than USD 10 000 in 2012 (see Figure 3.10 in Chapter 3 of this volume). Whole genome sequencing programmes involve many terabytes of data. The Cancer Genome Atlas (TCGA) that uses next generation sequencing technology is expected, for example, to generate approximately 2.5 petabytes (PB) of data. The enormous growth of genomic and other biomedical data are at a scale that makes traditional centralised approaches to data management and analysis impractical. Centralised approaches have become untenable for individual laboratories and most small to medium research organisations due to the high cost of data storage, transmission, and analysis.

Box 8.2. **Advances in genetic sequencing (cont.)**

In this big data landscape, new models of global scientific collaborations are emerging that rely on shared e-infrastructures, cloud-based consortia and advanced computational capacities. The aim is to extract knowledge from these large streams of data in the most effective way possible, through global collaboration, open science, and bringing "computing to the data". An example is the International Cancer Genome Consortium (ICGC), a multidisciplinary, international collaborative effort on the part of nine countries to systematically and comprehensively characterise somatic mutations in over 24 000 tumour genomes from 50 different cancer types and subtypes, comparing tumour and normal tissues.

Since its launch in 2008, the ICGC has generated over 250 terabytes of data, adopting federated data architecture to address the data management needs. The scalability of the system is improved by having each member institution store and process data locally; the data federation software then presents these separate sources as a single access point for remote data access.

Source: NIH (2014), "DNA Sequencing Costs", National Human Genome Research Institute, National Institutes of Health, www.genome.gov/sequencingcosts/.

Systems biology to model complex molecular mechanisms

Network and systems biology strategies today offer a powerful means to explore the complex molecular mechanisms underlying many diseases (Chen, Shen and Sivachenko, 2006; Liu et al., 2006; APA, 2006). Research efforts are now increasingly directed to better understanding the interactions between cellular components (proteins, genes, metabolites and so on) (Vidal, Cusick and Barabasi, 2011; Barabasi, Gulbahce and Loscalzo, 2011).

In humans, the potential complexity of the resulting networks – the human interactome – is daunting, with the number of cellular components that serve as the nodes of the interactome easily exceeding 100 000. The number of functionally relevant interactions between the components of this network is expected to be much larger. With so much data available, the challenge is to integrate that information into a single meaningful interaction network.

The highly interconnected nature of the interactome also means that at the molecular level, it is difficult to view diseases as being consistently independent of one another. Indeed, different disease molecular mechanisms can overlap, so that perturbations caused by one disease can affect other diseases (Barabasi, Gulbahce and Loscalzo, 2011).

The systematic mapping of such networks has therefore culminated recently in the concept of the "diseasome": disease maps whose nodes are diseases and whose links represent various molecular relationships among the disease-associated cellular components.

Progress towards a reliable network-based approach – however promising – is currently limited by the incompleteness of the available interactome maps, and the need for powerful visualisation tools as well as statistical methods that are reliable in the context of interconnected environments (Barabasi and Oltvai, 2004). With continued advancements in data analytics, systems biology has the potential to yield a much more nuanced understanding of disease processes, and a greater personalisation of treatments.

Big data for early detection of neurodegenerative disease

Conventional structural neuroimaging, such as computed tomography (CT) or magnetic resonance (MR), has long played a supportive role in diagnosing memory disorders, and is today recommended for the routine evaluation of Alzheimer's disease (AD). However, because structural changes may not be detected at visual inspection until late in the course of the disease, more contemporary structural imaging techniques have emerged that aid in detecting subtle changes not readily apparent on routine images obtained at a single time point. These include positron emission tomography (PET), single photon emission CT (SPECT), and functional magnetic resonance imaging (fMRI).

Functional magnetic resonance imaging in particular offers the promise of revolutionary new approaches to studying human cognitive processes, provided we can develop appropriate data analysis methods to make sense of the huge volumes of data. fMRI measures brain activity by detecting changes in blood flow and blood oxygen levels (the ratio of oxygenated haemoglobin to deoxygenated haemoglobin in the blood with respect to a control baseline), over time and at many individual locations within the brain. It is widely believed that the blood oxygen level is influenced by local neural activity, and hence it is generally taken as an indicator of that activity.

A twenty-minute fMRI session with a single human subject produces a series of three dimensional brain images, each containing approximately 15 000 voxels, collected once per second, yielding tens of millions of data observations. Each voxel contains hundreds of thousands of neurons.

Accurate quantification of changes in regional brain volumes is time- and labour-intensive. If this limitation of fMRI-based methods can be solved via automation of scan analysis, such methods are almost certain to become useful tools for the early detection and monitoring of Alzheimer's and other neurodegenerative diseases in patients.

The Quantitative Imaging Network (QIN), driven by the US National Cancer Institute, grows out of this need to improve translational and clinical research in imaging sciences and technology. The network is designed to promote research in and development of quantitative imaging methods for the measurement of tumour response to therapies in clinical trial settings, with the overall goal of facilitating clinical decision making. Projects include the appropriate development and adaptation/implementation of quantitative imaging methods, imaging protocols, and software solutions/tools (using existing commercial imaging platforms and instrumentation), and application of these methods in current and planned clinical therapy trials.

Sensor-based systems to monitor behavioural changes

Sensor-based systems can also be leveraged to provide clues on emerging physical and mental health problems. Ubiquitous sensors have, for example, an increasingly important role to play in integrating a novel and less-biased window of cognitive and behavioural monitoring of older patients. These systems can also provide assistance in increasing the independence and security of people who have problems of memory, planning, and carrying out tasks in everyday life.

For individuals with chronic conditions, unobtrusive home-based monitoring can result in better patient outcomes by allowing the physician to verify compliance with pharmaceutical regimens and activity-level guidelines to better understand the range of variation of patient outcomes.

For older patients, sensor-based devices can also be utilised to monitor falls and near-falls, physical activity, socialisation, and even overall mobility. For example, wearable fall detectors that include accelerometers are a good example of information technology for assisted living at home (Brown, 2005). In most of these systems, a periodic report from the sensors is sent via wireless communication to a local base station.

The biggest hurdle to overcome to make these approaches useful is the development of efficient and user-friendly data-flow processing and effective conversion of the sensor events into clinically actionable knowledge. Context awareness imposes significant demands on the knowledge maintained by these systems.

Progress will thus depend on the development of robust algorithms and computational models that can fuse and derive meaning from the diverse sets of information. Key factors influencing scalability include: i) seamless integration and interoperability of the technology; ii) reliability of message capture, translation, and delivery to health care professionals and the amount of information transmitted per patient; iii) frequency of monitoring and transmission, and context awareness.

Using social media to research population and public health

Web- and mobile-based applications of social media are emerging as useful new approaches for the dissemination and collection of health and lifestyle information. They can reach a broad audience in a very short period of time; they are easy and affordable to access and use; and they cater to a wide variety of people.

Online social communities, for example, provide a vehicle for individuals with chronic diseases to share information on therapies and disease progression. Participants contribute personal stories that provide learning experiences for other participants who may be contending with a similar health problem. Some online communities are moderated by health care professionals who can offer expert advice via message board posts or synchronous chat sessions.

There is growing recognition that online communities not only provide a place for members to support each other, but also contain knowledge that can be mined for public health research, monitoring, and other health-related activities. By harnessing the power of global, widely disseminated user-generated content, social media is increasingly proving itself an important communication platform on health and disease, serving as an opportunity to collect data on patients' experiences to guide policy and communication planning. At the same time, analytical uses of social networking data must protect the privacy of data subjects. A lack of adequate methods to respect privacy in the use of this data can be a barrier to that use.

For example, the social network PatientsLikeMe developed a lithium-specific global data collection process to capture information about individuals suffering from amyotrophic lateral sclerosis (ALS) who were registered with the network and who began taking the drug off label via their physician (Wicks et al., 2011; CDC, 2009).

ALS is a chronic condition for which neither randomised trials nor nonrandomised clinical studies have yet provided an effective therapy. It is a rapidly fatal neurodegenerative disease causing progressive weakness and muscle atrophy; median survival from symptom onset is 2-5 years. In 2008, a small study suggested that lithium carbonate slowed ALS. Once that study was published, hundreds of ALS patients on PatientsLikeMe began taking the drug, and a few used freely available tools such as Google spread sheets to "crowdsource"[20] their own study. In response, PatientsLikeMe

upgraded its tools and developed new analytical techniques to evaluate whether lithium was effective (Wicks et al., 2011).

The social network also regularly imports the complete data set from ClinicalTrials.gov to let its membership know (free of charge) about the 30 000+ active trials for which they may be eligible. Government agencies are also using social networks to engage the public – for example, during product recalls and pandemic preparations (e.g. in the H1N1 flu pandemic) and as resource for investigating drug-related activity such as off-label use, side effects, product safety, and patient opinions.

Twitter (www.twitter.com) is also emerging as a suitable platform for this purpose. Twitter allows users to send and read short text-based messages limited to 140 characters, which contain a wealth of data. Mining these data provides an instantaneous snapshot of the public's opinions and health-related behavioural or other responses. Longitudinal tracking allows identification of changes in opinions or responses. In addition to quantitative analysis, twitter also permits qualitative exploration of likely reasons why sudden changes have occurred (e.g. a widely read news report), and may indicate what is holding the public's attention.

Twitter content has been studied recently to track flu epidemics (Chew and Eysenbach, 2010) to assess public misunderstandings surrounding antibiotic use (Scanfeld, Scanfeld and Larson, 2010; Signorini, Segre and Polgreen, 2011 ; Sadilek, Kautz and Silenzio, 2012), and more recently to gain insights on how online users share information about dementia and the type of information shared (Robillard et al., 2013).

There is also a large amount of literature proposing methods to extract useful information from online data-generated searches each day, including through Yahoo! and Google (Eysenbach, 2006). A Centers for Disease Control study conducted with Yahoo! in 2005 suggested that Internet searches for specific cancers correlated with their estimated incidence, estimated mortality, and volume of related news coverage. The authors concluded that "media coverage and prevalence appeared to play a powerful role in prompting online searches for cancer information" (Cooper et al., 2005).

Although current Internet search query data are no substitute for timely local clinical and laboratory surveillance, recent studies indicate that the intensity of certain web queries on influenza and influenza-like illness follows the same pattern as the laboratory and sentinel reports for influenza, and that they can be used as additional input data for estimation models (Hulth, Rydevik and Linde, 2009; Eysenbach, 2007).

In November 2008, Google Flu Trends was launched as an open tool for influenza surveillance in the United States. Engineered as a system for early detection and daily monitoring of the intensity of seasonal influenza epidemics, Google Flu Trends uses Internet search data and a proprietary algorithm to provide a surrogate measure of influenza-like illness in the population (Olson et al., 2013; Yin, 2012). The algorithms may still need to be refined however, as the journal *Nature* (2013a) reported in February 2013 that Google's Flu Trends data was significantly overestimating the number of influenza cases. Some researchers suggested that widespread news coverage led to spikes in influenza-related searches by people who were not ill (*Nature*, 2013a).

Crowdsourcing health care innovation

Crowdsourcing is emerging as a means to allow science to be conducted at scales of magnitude greater than before. It involves capitalising on the Internet and large groups of people, particularly via online Web 2.0 communities, to harvest "collective intelligence"

and accomplish tasks that might have traditionally been given to small research groups. Crowdsourcing is for example successfully being used by foundations and the public and private sectors for health research purposes such as to understand protein structure prediction and design.

Crowdsourcing can process data quickly and on unprecedented scales and with better quality control than any individual or small research group can attain, given the large number of participants. Crowdsourcing therefore has cost and speed benefits; it may allow science to be conducted at scales of magnitude greater than before (thousands of research participants recruited in months versus years) and huge numbers of data points, the potential for new discoveries in the patterns of large data sets, and the possibility of near real-time testing and application of new medical findings.

The success rate of crowdsourced innovation challenges is quite high, in some cases up to 40% – which is remarkable, especially since many "challenges" are generally put out on the web because they are, by definition, beyond the problem-solving ability of the organisation or the individuals posting them. Given its open, informal structure, crowdsourcing is inherently cross-disciplinary. In some cases, even gifted amateurs and people without direct experience with the problem provide valuable insights and solutions. However, at present there are several important unresolved ethical and legal issues that limit the use of crowdsourcing in health research (see also Chapter 5 of this volume):

1. Crowdsourcing may accelerate the sharing of information, but careful attention must be paid to policy regarding privacy, security, data stewardship and personal control. Rapid developments in this area have outpaced regulatory frameworks, raising a number of concerns that range from the potential of modest risks to the privacy of participating individuals and to the quality assurance of the large streams of data generated to severe safety risks.
2. Second, a range of new partnerships is emerging around these applications. There is a need to better understand this rapidly evolving ecosystem – the business models, the market potential and the related governance frameworks. This should be combined with the development of robust metrics for measurement and evaluation.

Despite these challenges crowdsourcing is increasingly being used by public and private sector to address complex and challenging problems. In 2011, 1 920 000 results were returned for a Google search of the terms crowdsourcing & health, linked with an ampersand; in 2010 and 2009 the figure was 669 000 and 318 000, respectively. In January 2012, the term "crowdsourcing" in a PubMed search yielded 16 publications, 13 of which were published in 2011 (Swan, 2012a).

InnoCentive, one of the first companies to crowdsource in the chemical and biological sciences, today has more than 300 000 registered "solvers", who stand to gain rewards of between USD 5 000 and USD 1 million if their solution works. Key to the success of InnoCentive's crowdsourcing has been: i) a governance structure carefully designed to protect intellectual property from both the Seeker and the Solver perspective; ii) reduced barriers to participation so that the challenge scales quickly; and iii) global reach, increasing the likelihood of solutions coming from very unexpected directions.

An illustration of public sector uses of crowdsourcing for health is the "Investing in Innovation (i2)" initiative of the US Office of the National Coordinator for Health Information Technology. The i2 program arranges challenge competitions to spur innovation in the developer community. These challenges, in which the winners are

awarded prizes, enable the sector to reach out to developers who have expertise in different fields: although they do not necessarily work in health IT, they can apply their knowledge from elsewhere to health IT issues.

An ongoing challenge, developed in partnership with the US National Cancer Institute, asks developers to come up with tools and applications for cancer survivors. Applications arising from the ONC challenge programme are already widely used – including Humetrix's iBlueButton, which enables patients to access and share their health records.

Citizen science initiatives are also growing in importance, internationally and in the bioscience field; often they are associated with crowdsourced challenges. Citizen science generally refers to a network of people, many of whom may have no specific scientific training, performing research-related tasks, such as recording specific observations over time to reveal patterns and trends. The approach leverages what Shirky (2010) called "cognitive surplus", referring to the vast amount of time that people collectively spend on activities such as watching TV.

Citizen science projects often involve non-professionals taking part in one or more of the following:

- crowdsourcing
- mass participation
- data collection
- data analysis.

Zooniverse, for example, works with researchers to design sites that present their data in a format that will permit the crowd help them achieve their objectives.[21] Zooniverse has a community of over 850 000 people, who have taken part in more than 20 citizen science projects over the years. These initiatives support a form of "scientific democracy", where data can be shared among and utilised by investigators in public and private sectors, policy makers, and the public.

Foldit is another popular online citizen science initiative, in which individuals are scored on noting changes in protein structure. The game records the structure and the moves that the players make; scientists can capture the data that are then used to improve the game in every aspect, from the quality of the scientific results that are returned to how long people play the introductory levels that teach the game. The whole game is like an ongoing, continuous experiment.[22] Foldit was successfully used to remodel the backbone of a computationally designed enzyme that catalyses the Diels-Alder reaction, bringing together two small molecules to form a particular kind of bond that the scientists were interested in making.[23] This catalysis can be useful in building other kinds of small molecules, such as drugs and chemicals. Scientists went back and forth with the players, and in the end designed an enzyme that was about 20 times more efficient in catalysing the reaction than the one the scientists had begun with. Tapping into the vast cognitive surplus online and incorporating crowdsourcing in research, to both enlist the public's help and engage public interest, holds tremendous promise for accelerating health innovation.

DIYgenomics, established in 2010, leverages crowdsourcing and citizen science in order to produce preventive medicine. At present the focus of DIYgenomics is on linking genetic mutation with phenotypic evidence and personalised intervention in a wide range of studies. DIYgenomics has vitamin deficiency studies under way to investigate two

possibilities.[24] The first is that one or more genetic polymorphisms (e.g. mutations) may lead to current blood marker levels that are already out of bounds per recommended levels; the second is that simple vitamin supplementation may be able to restore blood markers to recommended ranges. DIYgenomics also has a study examining how genetic variants may be related to dopamine processing in the brain, and how this may impact the processing of memories. DIYgenomics moreover hosts a longitudinal ageing study aiming to establish personal baseline norms for 50 blood markers and their potential correspondence to 1 000 gene variants associated with ageing, and experimenting with personalised intervention. The study provides an opportunity to apply the dozens of genome-wide association studies (GWAS) that relate to general and specific conditions of ageing in a comprehensive preventive medicine approach. Genomic data are linked with corresponding measures of phenotypic biomarkers and interventions. The participants' tasks are to complete an annual blood test (a comprehensive panel of approximately 50 markers available through DirectLabs [USD 79] or another source) and, if willing, share the data with the cohort and self-experiment with relevant interventions. One thousand genetic variants are reviewed that have been linked to a variety of ageing conditions (Swan, 2012a, 2012b).

Funding for crowdsourced and citizen science health studies comes from a variety of sources, including foundations, academia, the private sector and patient advocacy groups. "Next-generation research foundations" are finding crowdsourced cohorts suitable for their studies. This is especially true of foundations that are focused on genomics and personalised medicine – such as Sage Bionetworks, a non-profit research organisation based in Seattle. That organisation's Alzheimer's Big Data Challenges, the Global CEO Initiative on Alzheimer's Disease (CEOi), and IBM's Dream Team are examples of bio-informatics crowdsourcing aimed at identifying the best multivariate set of predictors of cognitive decline and of neuro-protective genotypes.[25]

The many concepts underlying crowdsourcing have been around for some time. But as the proportion of the world's population with access to the Internet climbs, the potential for crowdsourcing to both generate data and help interpret data will grow. At the same time there are challenges that must be overcome to allow the generalised use of crowdsourcing to improve health research and health care. These need to be addressed through supportive policies and environments. The next section outlines the factors that support data-driven research and care.

8.5. Critical success factors and policy priorities

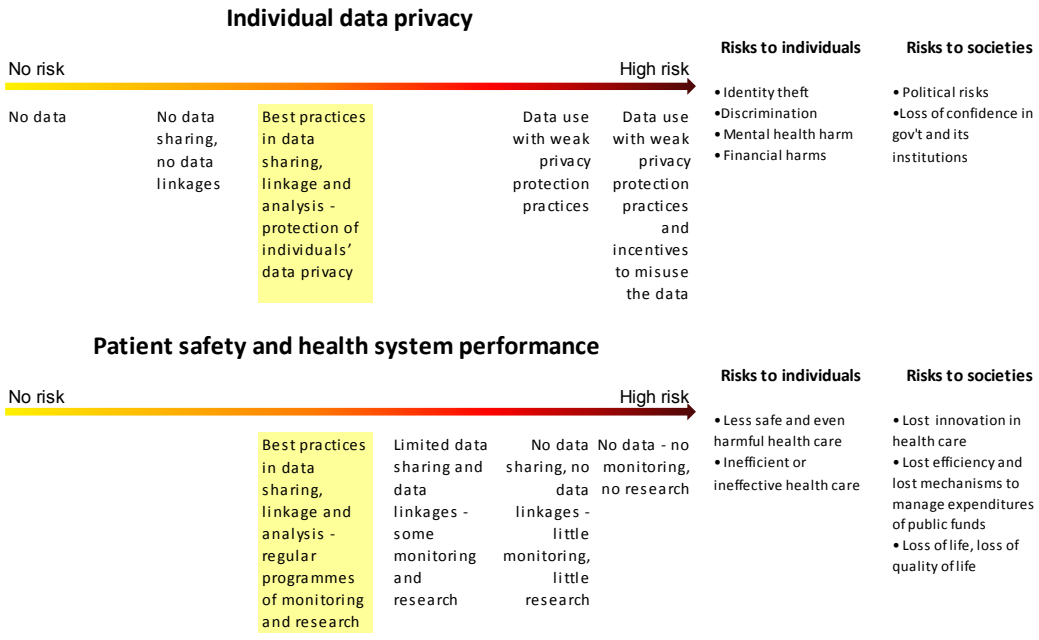
The most critical success factor for governments to realise value from investments in health data is data governance that is proactive, engaged and comprehensive. Such governance must be multi-sectoral, as the needs and conditions that would lead to a data-driven health sector reach beyond the control of health ministries. This section reviews priorities for policy action.

Minimising risks to the data subject's privacy

Collection and use of personal health data present a number of important risks to the privacy of individuals (Figure 8.3). These risks relate to the potential harm to individuals that could result from the misuse of their personal information. Losses to individuals can be severe and can include financial and psychosocial harm. Detriment to individuals can also produce a loss of public confidence in government and its institutions. Yet, there are equally significant risks to individuals and to societies when health information assets are

not developed, are unused, or are very difficult to use. These include a lack of evidence to detect and correct inefficient, ineffective and even harmful health care, and lost opportunities for research and innovation to improve health and health care outcomes.

Figure 8.3. **Risks associated with the collection and use of personal health data**



Source: OECD, 2013e.

Significant differences in approaches to the protection of data subjects' privacy among OECD countries have resulted in some countries advancing the generation of health data and its use for research and statistical purposes, and others restricting data collection, sharing and use. These differences are significant and can be attributed to differences in risk-benefit evaluations.

Many OECD countries report legislative barriers to the use of personal health data, including enabling data linkages and developing databases from electronic health records. Some of the countries with less developed information infrastructure have decentralised administration of health systems, and have not reached a consensus within the country of how the levels of government could work together. A principal challenge for some countries is the lack of clarity on how to translate into practice legislation concerning the protection of data privacy, including informed consent at the national and sub-national levels. This includes the legality of data sharing among public authorities and that of providing access to personal health data for research.

This complexity extends to multi-country data-sharing initiatives; the result is that such initiatives remain rare, challenged by concerns regarding differences in data privacy protection laws and whether shared data will be adequately protected in the receiving country. European countries have made the most progress among OECD members, having recognised in law that foreign entities can apply and be approved for access to data when the legal protection of information privacy in the foreign country adequately matches that of the home country. However, lack of resources to evaluate the adequacy of foreign laws continues to pose barriers to data sharing between European countries and other countries.

The need for new consent models

An important requirement prior to personal data collection for health research is to obtain patient consent. Explicit consent has become the pillar for protecting autonomy in research involving human subjects. The requirement for consent is underpinned by ethical principles of respect for persons. Consent is also the basis for data protection and privacy laws in most countries. Within the medical/scientific field, informed consent generally presumes the ability to indicate clearly to the participant the use and purpose of the particular research activity. While this is feasible for purpose-specific research, the new forms of biomedical research – and the ease with which multiple data from diverse sources can now be collected, stored, analysed and shared in greater volumes than ever before – renders provision of this type of information particularly difficult.

In the case of biobanks, for example, where there are multiple researchers and research projects, it is difficult to obtain explicit consent for all future research uses at the time of research recruitment, as is required in the original formulations of the Declaration of Helsinki (World Medical Association, 1964). The declaration states that use for research purposes different from the original would require re-contacting large population groups to obtain a new consent, which is often impossible or impracticable. Re-consenting is costly and time-consuming, and difficulty in locating people can result in high dropout rates. New approaches are clearly needed to meet ethical and legal requirements for consent and to accommodate the changes in data use and research practices.

A tiered or step-by-step consent approach has recently been adopted at Imperial College in the Chariot Register, a recently established cohort of over 20 000 healthy volunteers for the prevention of dementia and other age-related neurodegenerative diseases. Participants initially consented to be approached for individual observational or interventional studies, and were then offered a menu of options pertaining to such research uses, requests to re-consent, interest in returning results, etc. Other approaches recently proposed in the scientific literature include "adaptive" or "dynamic" models of consent forms, whereby (following the initial general consent) participants would be asked to re-consent for any "new" direction of travel/use of their data, potentially using web-based communication tools. This approach is dynamic because it allows interactions over time; it enables participants to consent to new projects or to alter their consent choices in real time as their circumstances change, and to have confidence that these changed choices will take effect (Kaye et al., 2014).

A key role for data custodians

Personal health data sets are in the custody of multiple organisations within countries, and legal frameworks and internal policies must be aligned to permit secure data sharing if big health data assets are to be brought together for research and statistics. Countries reported encountering difficulties in negotiating data-sharing arrangements among national organisations, with negotiations either unsuccessful or taking years to conclude (OECD, 2013e).

Countries that have centralised processing of personal health data within a single or a small number of organisations have the advantage of avoiding the need for complex data-sharing negotiations, as well as gaining efficiencies in data processing and data security protections.

These data custodians also play a central role in balancing data privacy protection and the use of data for monitoring and research, as they are responsible for the collection, processing, analysis and dissemination of personal health data. In many countries, data custodians are also responsible for vetting project proposals for the use of data from government and private entities; for maintaining the technical capacity to undertake data linkages; for maintaining a technical capacity for data de-identification; for providing data access modalities to internal and external researchers; and for ensuring that through all of their activities, the legal requirements for data security and data privacy protection are respected.

Development of privacy-enhancing technologies

Advances in techniques to ensure privacy through the design and development of privacy-enhancing technologies provide additional avenues to meet both health care data use and privacy protection needs (OECD, 2013e).

The practice of data de-identification and data pseudonymisation are widely used across countries, particularly before data are made available for research and analysis. While de-identification involves the removal of key patient identifying information, such as names, patient numbers, exact addresses and key dates, pseudonymisation replaces key patient identifiers with a meaningless code that can, for approved purposes, allow re-identification. An example of an approved purpose could be to conduct a new approved data linkage study with the same data set in the future. With this approach, a trusted party is usually employed to guard the key that enables data re-identification.

Data de-identification techniques, however, rarely remove all risk that a data set could be manipulated or combined with other data to rediscover the identity of data subjects. Importantly, some countries have developed data governance mechanisms that provide added security to protect de-identified data. These include independent review bodies that evaluate data use proposals for public benefits and adequacy of data security; contractual agreements that bind data receivers to required data security and disclosure practices; and security audits and follow-up mechanisms to ensure compliance with contractual obligations.

A few countries are pioneering alternatives to sharing de-identified data in order to further minimise data security risks. These include supervised research data centres, where authorised researchers analyse data within a physically secure location; and secure remote data access services, where authorised researchers enter a secure portal; there they can analyse data but cannot extract or otherwise remove data from the system (OECD, 2013e).

Engaging stakeholders

In an OECD 2012 survey of 25 countries, 13 reported having involved groups of stakeholders in their efforts to govern the development and implementation of their national electronic health record system, either through the groups' representation within the governing body or through consultation, or both. The groups included, for example, clinicians, pharmacists, professional associations, patients, insurers, IT professionals, lawyers and policy makers (OECD, 2013e). Engaging with all interested stakeholders would appear to be the best strategy for ensuring that all voices are heard and a consensus is reached on data use that respects privacy. Further, a public communication strategy that is open and transparent would go a long way toward demystifying data, opening data for monitoring and research, and generating positive public discourse about data risks and

utilities. Ideally, the strategy would enable all concerned stakeholders to know what data are being collected; how they are being used; how and with whom to apply for access to them; the conditions of approval; data security requirements; and details of the research projects that are approved.

Even with strong communication with the public; a high degree of transparency regarding data uses; and strong governance to safeguard patients' privacy in law and in practice, attention is needed to ensure that individuals who wish to restrict or withdraw their data from their contribution to research and statistics can reasonably do so. Strategies to enable individuals to exercise control over the use of their personal health data must be workable at the scale of population-level data collection, and in circumstances where there is a high volume of data and a high number of data use requests. Emerging techniques reported to the OECD in 2013 include the use of Internet patient portals, to request patient consent to data uses or to enable patients to opt out of data use. The portals provide information about data use that is broad enough to capture future uses for research and statistics that cannot be specified today, while being narrow enough to allow patients to fully understand the circumstances under which their data could be approved for use and how they would be protected throughout that use.

Promoting open data and data commons

Governments in several OECD countries have been engaged in initiatives to increase the openness and transparency of government data, including health data. Advocates for greater openness and transparency link the availability of government data and information to more socially inclusive service delivery; to participatory democracy; and to economic stimulation from the development of new products and services (Chapter 10 of this volume). In 2011, eight governments (Brazil, Indonesia, Mexico, Norway, the Philippines, South Africa, the United Kingdom, and the United States) founded the Open Government Partnership, and announced their country's action plan towards open government health data (UK Department of Health, 2012). This partnership has since grown dramatically, to 47 additional governments.

In 2013, of the 20 countries participating in the OECD Health Care Quality Indicators (HCQI) survey, 12 indicated that their country is planning a policy or programme to promote open government health data. For most countries, this effort is part of a whole-of-government initiative to provide citizens with a single entry point to government statistics, including health in the form of a web portal. In a few countries, this also includes developing mechanisms for citizens to more easily retrieve their own personal data.

The Health Data Initiative in the United States aims to increase data availability and transparency in order to improve community health. In particular, the initiative provides access to a broad range of health data at the local, state and national level, disaggregated by socio-economic characteristics, standardised and documented to enable ease of use.[26] The US initiative also involves working with clinicians, information technology professionals, policy makers and citizens to develop software applications and tools that turn data into actionable information, be they smartphone apps, interactive maps, indicators, social networking sites or games, etc. (DHHS, 2013).

In the United Kingdom, England launched a ten-year strategy to improve the National Health Service, public health and social care (UK Department of Health, 2012). An open government health data programme is an integral part of this strategy; it includes routinely releasing public service data sets in health to the public; providing health

service users with access to their own data; gathering and publicising health services user satisfaction and experiences data; engaging with data users to drive social and economic growth; and working to continuously improve data quality. The open government health data programme also connects with the overall strategy of the United Kingdom to promote open government data (United Kingdom, 2012).

The Australian government also has an open government data initiative with the inclusion of some government health data, as do Canada and Italy.[27] With its Digital Agenda for Europe, the European Commission is also promoting greater openness and reuse of government data, some of which is related to the health care system.[28] This project includes establishing a website to disseminate data held by the European Commission, and work toward a pan-European portal for all data from the EU, as well as from national governments and regional and local governments in Europe.

Building a new generation of health data scientists

Data scientists with good communications skills, and clinicians and other health professionals at ease with numbers and computing, can work together to produce remarkable results. Developed from a health data initiative of the Institutes of Medicine and the Department of Health and Social Services in the United States, a new event called Health Data Palooza was held in 2012 and 2013.[29] This event brings a diverse set of stakeholders together to discuss obtaining value from data. One of its key elements is a 48-hour competition where teams of clinicians and IT professionals compete for prizes by developing an app, tool or product from analysis of US Medicare databases. At the same event, start-up and established companies can showcase products leveraging information value from health data. In 2013, 80 companies showcased new products.[30] In England, events called NHS (National Health Service) Hack Days bring together clinicians, programmers and website designers. Some of their products that are now in use include Cell Countr, an app that provides haematology counts, and PatientList, an app that provides clinical task lists (Lewis et al., 2013)

England plans to engage clinicians directly in developing tools to exploit the power of big data to improve the NHS (Lewis et al., 2013) There are plans to involve medical professionals, patients and NHS managers in the development of applications and tools by training them how to write computer code; the initiative is called Code-4-Health. The thinking is that clinicians and managers with some knowledge of code will be better equipped to work with IT professionals toward developing useful tools, and able to unleash their creativity and innovative ideas. It will also avoid the old style of computer applications development, where a detailed specification is given to a programmer who works alone to try to develop code to meet it, often to the dissatisfaction of the end user.

Nonetheless, a scarcity of data scientists needs to be avoided through education and training initiatives, and such training needs to be responsive to the necessity of data scientists having the teamwork and collaboration skills to partner with health care professionals (Davenport and Patil, 2012). Adapting education and training programmes for other health care professions to ensure at least a minimum degree of skill development in statistics and programming is another worthy objective, both to build a generation of clinical data scientists and to increase appreciation for data and high-quality record keeping within the health care professional community (see further discussion on the skills implications of DDI in Chapter 6 of this volume).

Financial sustainability of big data projects

Big data is a costly activity. As an illustration, the cost of implementing the Canadian longitudinal study on ageing – which includes the first wave of data collection, follow-up on the initial cohorts, and management – was estimated at CAD 23.5 million. Generation of big data (e.g. imaging, microarray, phenotypi data etc.) can include costly processes, requiring expensive consumables as well as specialised equipment and personnel for their generation. If for financial reasons these networks or databases are unable to perform their tasks under conditions that meet the requirements of scientific research, scientists will see valuable information either lost or transferred into a strictly for-profit environment.

There is no magic bullet today with respect to the options or strategies required to achieve the long-term financial sustainability of big data projects. Financial sustainability is a critical issue for all big data initiatives – even those that are relatively more mature and directly funded by public sources.

As ongoing financial support is uncertain, large data networks very often must seek out multi-source financing – for example, by charging fees to those who want to gain access to a specific data set and associated database. Varying fee structures can be applied for access, depending on the nature of the data, its status and use. Another model that appears to have great potential for the prolonged financial sustainability of big data projects is the public-private partnership.

Increasing accessibility and sharing of existing data can be resisted due to mismatches in resources and incentives. Data collectors may not have sufficient resources to meet data access requests. Policies are needed that consider the bigger picture regarding the benefits of data use to improve innovation in treatment and care; they must provide incentive structures and resources that will enable key data holders to take part in making the necessary data available.

Setting standards to enhance interoperability

Standards and interoperability are central issues that must be addressed to advance big data in health care. While health care organisations have access to an ever-increasing number of information technology products, many of these systems cannot "talk" to each other and health information exchange remains a serious problem. If the systems cannot communicate, big data will not meet its potential in the health care system. Ensuring that electronic records can be transferred or shared among a patient's primary care physician and specialists is an issue that has yet to be addressed. This problem is common to all OECD countries, even those where deployment of EHRs has proved particularly successful.

In a networked environment, interoperability means common protocols defining the basic mechanisms by which users and resources negotiate, establish, manage and exploit data-sharing relationships. It means sharing not only data but anything that connects to the data's production and processing, including computing tools, applications, methods, software, metadata, workflows across different platforms and, as mentioned above, even communication.

There is, for example, a clear need to develop and promote international consensus standards for clinical information. The World Health Organisation has developed classification systems, which are essential for research and statistical reporting; however, these systems need to evolve to meet the needs of electronic clinical records and other

new forms of electronic data. Strategies to address a lack of standards for clinical terminology within electronic health records and to improve data quality and coverage include: laws or regulations requiring adoption and use of EHR systems that conform to standards; incentives and/or penalties relating to adoption and use; certification of software vendors; and quality auditing (OECD, 2013e; OECD, 2010). Many governments have set up specific bodies or agencies to co-ordinate standard-adoption activities, developing strategies at the national level.

Nonetheless, EHRs are expected to have important limitations compared with a clinical research data set designed to follow a strict research protocol with standardised definitions and rules to optimise statistical analysis. Simply put, the actors involved in inputting the data to EHR systems are primarily motivated by very different incentives – such as simplifying notes for their own use and saving time – from those of clinical scientists gathering data for their own research.

Technical solutions are on the horizon that could alleviate the burden of completing EHRs and address data quality problems. In particular, the advent of natural language processing (NLP) may enable health care providers to speak rather than to type; importantly, they may also require less personal knowledge of and competency with clinical terminology code sets (see Chapter 3 of this volume). This technology has not yet advanced sufficiently, however, for it to be widely adopted for use in EHR systems (Friedman, Rindflesch and Corn, 2013).

Analytic techniques are also emerging, to better cope with heterogeneity among genome experiment data, and to enable data to be pooled for systems-level research. These techniques include Bayesian integration, which enables prediction of the probability of an interaction between gene pairs and quantifies the contribution of each experiment to the prediction (Greene and Troyanskaya, 2012). Nonetheless, global standards for genomic data that would facilitate research are needed (*Nature*, 2013b).

The final area where standards are needed is in ensuring that observational health studies involving analysis of large databases follow scientifically sound methodologies. Certainly, the wealth of detail within electronic health records provides the ability to address challenges faced in the past due to an absence of information about important confounding factors. New methodological designs take this potential further and advance the science of comparative effectiveness research. These advances include sequential cohort studies, extensions of clinical trials with health record data, and modelling and trial simulation (Schneeweiss et al., 2011).

Providing incentives, investments, and grants

Options for stimulating the development and use of data abound. Eleven of twenty-five countries indicated in 2012 that they have introduced incentives, penalties or both to encourage health care providers to adopt EHR systems that conform to standards and use structured data. Seven countries have also introduced incentives or penalties to ensure that health care providers keep their electronic health records up to date. Eight countries have implemented or were implementing legislation or regulations requiring health care providers to adopt electronic health records and/or to conform to clinical terminology and interoperability standards.

The US National Institutes of Health recently launched the Big Data to Knowledge Initiative, providing grant funding to create centres of excellence in big data computing in the biomedical sciences. Objectives include developing new policies to encourage data

and software sharing; a catalogue of research data sets; data and metadata standards development; analytical software development; access to large-scale computing; engagement with users and software developers; increasing the number of computational and quantitative researchers; and strengthening the skills of existing biomedical professionals.[31] Further, as was previously noted, the USD 1.1 billion in federal stimulus funding in the United States has had a profound impact on the development of comparative effectiveness research. Over the next four years, Horizon 20/20 – the scientific granting programme of the European Union – intends to fund health research projects that involve the collection and processing of data for large populations, for long-term follow-up studies as well as studies to support the development of health data infrastructure (European Commission, 2015).

As was also noted earlier, the advent of prize competitions as an incentive is a new tool being leveraged in some countries to mobilise the business, clinical and research communities toward the use of health data to solve challenges – largely through the development of new computer tools and applications.

As decisions regarding investments in data infrastructure and analysis are taken, consideration should also be given to the balance between who is investing and who will receive the benefits from the investment. In situations where costs are borne by the state and benefits accrue to private sector businesses, compensatory mechanisms to share costs or to share profits should be considered.

In summary, the evidence discussed in this section strongly suggests that the most important prerequisite for advancement of DDI is proactive and comprehensive data governance. The next section summarises the key findings from this chapter and describes how work at an international level could help countries make further progress toward data-driven health research and care.

8.6. Key findings and policy conclusions

Data-driven innovation in the health sector is already taking place. This chapter has presented examples from many countries of how the huge volumes of clinical, genetic, behavioural and environmental data that can now be generated are being processed, analysed and integrated to support patient care, health system management and research. These advances have yielded extraordinary insights into the natural history of diseases and their diagnosis, prevention and treatment. There has also been a rapid increase in medical devices that produce streams of personal health data, as well as data provided by patients through personally controlled patient records and patient social networks – sources from which the research potential is beginning to emerge.

Data that encompass patients' care pathways – that is, that capture patient treatment flows through the medical system based on diagnoses, procedures, drug events and the outcomes and costs of those pathways – have tremendous potential to uncover medical errors, avoid adverse drug reactions, detect fraud, improve adherence to clinical guidelines and develop effective treatments. Care pathway data can support both scientific knowledge generation and intelligence for health care policy making and management. They also increase opportunities for synergy and feedback between R&D activities and policy and management activities. Electronic health records are an important resource for data on clinical care, and can contribute to examining patient pathways and their outcomes.

The most critical success factor for governments to realise value from investments in health data is proactive, engaged and comprehensive data governance. Elements of data governance proposed include strategic planning; ensuring legislative and regulatory requirements that support planning; introducing effective data privacy and security practices; engaging all stakeholders in planning and governance; developing a new generation of data scientists; promoting global co-operation; setting standards for data quality; and providing financial stimulus toward data development and use.

Countries that plan now for how they will harness the value of personal health data in a secure and regulatory-compliant fashion will have the opportunity to reap the considerable benefits of health care innovation, with the ensuing advantages of both high-performing health care systems and growth in health care innovation.

At an international level, the OECD will continue to support countries in strengthening their health information infrastructure and analytical capacity. Key opportunities for future international co-operation include:

- exploring technical and legal modalities that enable global sharing, linking and analysis of data for health research and health system performance improvement
- continuing support of countries in their quest to implement quality indicators to assess the performance of health services and systems through the OECD Health Care Quality Indicators (HCQI) programme
- ongoing monitoring of country progress in the development and use of health data, and exploring country skill development and technical infrastructure capacities to support data analytics
- reporting on emerging forms of health and wellness data, such as genetic data, medical device data and social media data
- reporting on government investments in data infrastructure and big data research in health care and on any cost-sharing or profit-sharing arrangements with the private sector that have been introduced to finance this work
- supporting the development of international coding standards for key elements of health data systems in collaboration with WHO
- exchanging information about how laws protecting privacy are implemented in this area and promoting approaches that effectively protect privacy while enabling the use of data for health research and health system performance improvement
- working with countries to identify good practices in data deposition, access, exchange and linkage to advance dementia and neurodegenerative disease research (OECD, 2013d).

Notes

1 The G8 Dementia Summit Declaration, released on 11 December 2013, is available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/265869/2901668_G8_DementiaSummitDeclaration_acc.pdf, accessed 7 May 2015. See also http://dementiachallenge.dh.gov.uk/category/g8-dementia-summit, accessed 7 May 2015.

2 The Global Public Health Intelligence Network (GPHIN), developed by Health Canada in collaboration with WHO, is a secure Internet-based multilingual early-warning tool that continuously searches global media sources such as news wires and websites to identify information about disease outbreaks and other events of potential international public health concern. See www.who.int/csr/alertresponse/epidemicintelligence/en/, accessed 7 May 2015.

3 HealthMap, developed at Boston Children's Hospital in 2006, uses online informal sources for disease outbreak monitoring and real-time surveillance of emerging public health threats. See www.healthmap.org/site/about, accessed 7 May 2015.

4 Information on the i2b2 initiative is available at: www.i2b2.org, accessed 7 May 2015.

5 Information on the eMERGE network is available at: https://emerge.mc.vanderbilt.edu, accessed 7 May 2015.

6 Information on the Kaiser RPGEH Program is available at: www.dor.kaiser.org/external/dorexternal/rpgeh, accessed 7 May 2015.

7 Information on the Million Veteran Program is available at: www.research.va.gov/mvp/default.cfm, accessed 7 May 2015.

8 Information on STRIDE is available at: https://clinicalinformatics.stanford.edu/research/stride.html, accessed 7 May 2015.

9 Information on the i4health network is available at: www.i4health.eu, accessed 7 May 2015.

10 Information on EMIF is available at: www.imi.europa.eu/content/emif, accessed 7 May 2015.

11 Information on the European Translational Information and Knowledge Management Services (eTRIKS) is available at: www.etriks.org, accessed 7 May 2015.

12 Information on the Integrative Cancer Research Through Innovative Biomedical Infrastructures (INTEGRATE) is available at: www.fp7-integrate.eu, accessed 7 May 2015.

13 A Next-Generation, Secure Linked Data Medical Information Space For Semantically-Interconnecting Electronic Health Records and Clinical Trials Systems Advancing Patients Safety In Clinical Research (Linked2Safety). Information is available at: www.linked2safety-project.eu, accessed 7 May 2015.

14 Information on the Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies (SALUS) is available at: www.salusproject.eu, accessed 7 May 2015.

15 Information on the Translational Research and Patient Safety in Europe (TRANSFoRm) is available at: www.transformproject.eu, accessed 7 May 2015.

16 Information on the European Collaboration for Healthcare Optimisation is available at: www.echo-health.eu, accessed 14 September 2013.

17 Information on the CONCORD Programme is available at: www.lshtm.ac.uk/eph/ncde/cancersurvival/research/concord, accessed 26 January 2015.

18 Information on CancerLinQ is available at: www.asco.org/institute-quality/cancerlinq, accessed 7 May 2015.

19 Similar conclusions were reached in the 2014 EC Green Paper on mHealth: http://ec.europa.eu/digital-agenda/en/news/green-paper-mobile-health-mhealth, accessed 7 May 2015.

20 Crowdsourcing is "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community" (Merriam-Webster, 2014).

21 Information on Zooniverse is available at: www.zooniverse.org/ (accessed 26 January 2015).

22 In Foldit puzzles, for example, players are rewarded for solving clashes and voids, places where the protein is not consistent with known biochemical patterns. Players are able to build a hypothetical protein and see how it works in the game. The game's score is based on a proxy for how well the protein would work in a laboratory; whether it can catalyse some reaction that the scientists are interested in; or how well the protein sticks to some part of a virus – or even, in the case of the Symmetry puzzles, how well the protein sticks to itself. Solutions that are promising are then synthesised in the laboratory.

23 Information available at: www.nature.com/nbt/journal/v30/n2/full/nbt.2109.html, accessed 7 May 2015.

24 Information on these studies is available at: http://genomera.com/studies/vitamin-d-study, accessed 7 May 2015.

25 Information on Sage Bionetworks is available at: http://sagebase.org/about-2/, accessed 25 May 2015.

26 For information on HealthData.gov, see: www.healthdata.gov/ (accessed 30 September 2013).

27 For information on the initiative: in Australia, see http://data.gov.au/ (accessed 27 September 2013); in Canada, see http://data.gc.ca/eng; and in Italy, see www.dati.gov.it/ (accessed 27 September 2013).

28 For information on the Digital Agenda, see: http://ec.europa.eu/digital-agenda/en/news/commission-welcomes-member-states-endorsement-eu-open-data-rules, accessed 27 September 2013.

29 For information on the event's history, see: http://healthdatapalooza.org/history-of-the-health-datapalooza/, accessed 17 September 2013.

30 For details on that year's event, see: www.whitehouse.gov/blog/2013/06/07/health-datapalooza-iv-tops-huge-year-health-data-liberation-innovation, accessed 20 September 2013.

31 For information on the Big Data to Knowledge Initiative, see: http://bd2k.nih.gov/about_bd2k.html#areas, accessed 23 September 2013.

References

AHRQ (2013), "What is comparative effectiveness research?", Agency for Health Care Research and Quality, http://effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/, accessed 16 September 2013.

APA (2006), "Funding outlook for NCI: News from the experts", American Psychological Association, May, www.apa.org/science/about/psa/2006/05/nci.aspx.

Barabasi, A.-L., N. Gulbahce and J. Loscalzo (2011), "Network medicine: A network-based approach to human disease", *Nature Review Genetics*, Vol. 12, pp. 56-68.

Barabasi, A.-L. and Z.N. Oltvai (2004), "Network biology: Understanding the cell's functional organization", *Nature Reviews Genetics,* 5, pp. 101-113.

Brown, G. (2005), "An accelerometer based fall detector: Development, experimentation, and analysis", University of California at Berkeley, July, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.6988/

Canadian Institute for Health Information (2013), "Better information for improved health: A vision for health system use of data in Canada", in collaboration with Canada Health Infoway**,** www.cihi.ca/cihi-ext-portal/pdf/internet/hsu_vision_report_en, accessed 15 May 2015.

Cancer Research UK (2013), *International Cancer Benchmarking Partnership (ICBP),* ICBP Publications, www.cancerresearchuk.org/health-professional/early-diagnosis-activities/international-cancer-benchmarking-partnership-icbp, accessed 14 September 2013.

CDC (2009), "H1N1 web and social media metrics cumulative data report April 22, 2009 – December 31, 2009", Centers for Disease Control and Prevention, Division of eHealth Marketing, National Center for Health Marketing.

Chen, J.Y., C. Shen and A.Y. Sivachenko (2006), "Mining Alzheimer disease relevant proteins from integrated protein interactome data", Pacific Symposium on Biocomputing, November, pp. 367-78.

Chew, C. and G. Eysenbach (2010), "Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak", PLoS ONE, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014118, accessed 15 March 2015.

Cooper, C.P. et al. (2005), "Cancer Internet search activity on a major search engine, United States 2001-2003". *Journal of Medical Internet Research*, www.jmir.org/2005/3/e36/, accessed 15 March 2015.

Dentzer, S. (2013), "Rx for the 'Blockbuster Drug' of patient engagement", *Health Affairs,* February, http://dx.doi.org/10.1377/hlthaff.2013.0037, accessed 15 March 2015.

DHHS (2013), "About the Health Data Initiative", Department of Health and Human Services, 30 September, www.hhs.gov/open/initiatives/hdi/about.html, accessed 30 September 2013.

European Commission (2015), Research & Innovation Participant Portal, https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/index.html accessed 15 May 2015.

Eysenbach, G. (2007), "Five-year prospective study harvesting search and click data from Google 2004-2007", presented at AMIA Annual Fall Symposium, Chicago, www.jmir.org/2009/1/e11/.

Eysenbach, G. (2006), "Infodemiology: Tracking flu-related searches on the web for syndromic surveillance", AMIA Annual Symposium Proceedings, pp. 244-48.

FDA (2013), *Mini-Sentinel*, US Food and Drug Administration, http://mini-sentinel.org/ (accessed 13 September 2013).

Fernald, G.H. et al. (2011), "Bioinformatics challenges for personalized medicine", *Bioinformatics*, Vol. 27, No. 13, pp. 1741-48.

Friedman C., T.C. Rindflesch and M. Corn (2013), "Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine, *Journal of Biomedical Informatics*, Vol. 46, No. 5, pp. 765-73.

Greene, C.S. and O.G. Troyanskaya (2012), "Data-driven view of disease biology", *PLoS Computational Biology*, Chapter 2, Volume 8, No. 12, http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.10028161002816.

Grossmann, C., B. Powers and J.M. McGinnis (eds.) (2011), "Digital infrastructure for the learning health system: The foundation for continuous improvement in health and health care – Workshop Series Summary", National Academies Press, Washington DC.

Häkkinen, U. et al. (2013), "Health care performance comparison using a disease-based approach: The EuroHOPE Project", *Health Policy*, 13 May, http://dx.doi.org/10.1016/j.healthpol.2013.04.013.

Hulth, A., G. Rydevik and A. Linde (2009), "Web queries as a source for syndromic surveillance", PLoS ONE, Vol. 4, Issue 2, www.plosone.org/article/info:doi/10.1371/journal.pone.0004378.

ICES (2013), "At a glance: Evidence guiding health care", Institute for Clinical and Evaluative Sciences, www.ices.on.ca/webpage.cfm?site_id=1&org_id=70, accessed 13 September 2013.

IHME (2013), "Global burden of disease", Institute for Health Metrics and Evaluation, www.healthmetricsandevaluation.org/sites/default/files/policy_report/2011/GBD_Generating%20Evidence_Guiding%20Policy_GBD2010_results.pdf, accessed 7 May 2015.

Jensen, P.B., L.J. Jensen and S. Brunak (2012), "Mining electronic health records: Towards better research applications and clinical care", *Nature Reviews Genetics,* 13, pp. 395-405.

Kaye, J. et alia (2014), "Dynamic consent: A patient interface for twenty-first century research networks", *European Journal of Human Genetics*, http://dx.doi.org/10.1038/ejhg.2014.71, accessed 15 March 2015.

Larsson, S. et alia (2012), "Use of 13 disease registries in 5 countries demonstrates the potential to use outcome data to improve health care's value", *Health Affairs*, Vol. 31, No. 1, pp. 220-27.

Laurence, J. et al (2014), "Patient engagement: Four case studies that highlight the potential for improved health outcomes and reduced costs", *Health Affairs,* September, Vol. 33, No. 9, pp. 1627-34; http://dx.doi.org/doi:10.1377/hlthaff.2014.0375, accessed 15 March 2015.

Liu, B. et al. (2006), "Exploring candidate genes for human brain diseases from a brain-specific gene network", Biochemical and Biophysical Research Communications, Vol. 349, pp. 1308-14.

Merriam-Webster (2014), "Crowdsourcing", Merriam-Webster.com, Merriam-Webster, www.merriam-webster.com/dictionary/crowdsourcing, accessed 24 September 2014.

Nature (2013a), "When Google got flu wrong", 13 February, www.nature.com/news/when-google-got-flu-wrong-1.12413, accessed 24 September 2014.

Nature (2013b), "Geneticists push for global data sharing", 4 June, www.nature.com/news/geneticists-push-for-global-data-sharing-1.13133, accessed 17 September 2013.

Nazi, K.M. et al. (2013), "Evaluating patient access to electronic health records: Results from a survey of veterans, *Medical Care*, Vol. 51, pp. S52-S56, March, http://journals.lww.com/lww-medicalcare/Fulltext/2013/03001/Evaluating_Patient_Access_to_Electronic_Health.11, accessed 6 March 2015.

NHS England (2013), "Care episodes statistics: Technical specifications of the GP extract", National Health Service, May, www.england.nhs.uk/wp-content/uploads/2013/08/cd-ces-tech-spec.pdf, accessed 13 September 2013.

NIH (2014), "DNA sequencing costs", National Human Genome Research Institute, National Institutes of Health, www.genome.gov/sequencingcosts/, accessed 3 May 2015.

OECD (2014a), "Unleashing the power of big data for alzheimer's disease and dementia research: Main points of the OECD expert consultation on unlocking global collaboration to accelerate innovation for Alzheimer's disease and dementia", *OECD Digital Economy Papers*, No. 233, OECD Publishing, Paris.
DOI: http://dx.doi.org/10.1787/5jz73kvmvbwb-en.

OECD (2014b), Health Statistics 2014., www.oecd.org/els/health-systems/health-data.htm accessed 15 May 2015.

OECD (2013a), Health Statistics 2013, www.oecd.org/els/health-systems/health-data.htm, accessed 15 May 2015.

OECD (2013b), *Health at a Glance 2013: OECD Indicators*, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/health_glance-2013-en.

OECD (2013c), *ICTs and the Health Sector: Towards Smarter Health and Wellness Models*, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264202863-en.

OECD (2013d), "Unlocking global collaboration to accelerate innovation for Alzheimer's disease and dementia", OECD Publishing, Paris, www.oecd.org/sti/sci-tech/unlockingglobalcollaborationtoaccelerateinnovationforalzheimersdiseaseanddementia.htm.

OECD (2013e), "Strengthening health information infrastructure for health care quality governance: Good practices, new opportunities and data privacy protection challenges", *OECD Health Policy Studies*, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264193505-en.

OECD (2010), *Improving Health Sector Efficiency: The Role of Information and Communication Technologies*, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264084612-en.

Okamoto, E. et al. (2013), "Evaluation of the health check up and guidance program through linkage of health insurance claims", *Journal of the National Institute of Public Health,* Vol. 62, No. 1.

Olson, D.R. et al. (2013), Reassessing Google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales", *PLoS Computational Biology*, Vol. 9, Issue 10, www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003256.

PCORI (2013), "Important questions, meaningful answers", Patient-Centered Outcomes Research Institute, www.pcori.org/, accessed 17 October 2013.

PricewaterhouseCoopers (2012) Emerging MHealth: Paths for Growth

Robillard, J.M. et al. (2013), "Aging 2.0: Health information about dementia on Twitter", PLoS ONE, Vol. 8, Issue 7, www.plosone.org/article/info:doi/10.1371/journal.pone.0069861.

Roederer, M.W. (2009), "Cytochrome P450 enzymes and genotype-guided drug therapy". *Current Opinion in Molecular Therapeutics*, Vol. 11, No. 6, pp. 632-40.

Sadilek, A., H. Kautz and V. Silenzio (2012), "Predicting disease transmission from geo-tagged micro-blog data", Conference Proceedings from the 26th AAAI Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp. 136-42.

Scanfeld, D., V. Scanfeld and E.L. Larson (2010) "Dissemination of health information through social networks: Twitter and antibiotics", *American Journal of Infection Control* Vol. 38, pp. 182-88.

Schneeweiss, S. et al. (2011), "Assessing the comparative effectiveness of newly marketed medications: Methodological challenges and implications for drug development", *Clinical Pharmacology and Therapeutics*, Vol. 90, No. 6, pp. 777-90.

Shirky, C. (2010) *Cognitive Surplus: How Technology Makes Consumers into Collaborators*, Penguin Group.

Signorini, A., A.M. Segre and P.M. Polgreen (2011), "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic", PLoS ONE. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467, accessed 15 March 2015.

Swan, M. (2012a), "Crowdsourced health research studies: An important emerging complement to clinical trials in the public health research ecosystem", *Journal of Medical Internet Research*, Vol. 14, No. 2.

Swan M. (2012b) "Scaling crowdsourced health studies: The emergence of a new form of contract research organization", *Personalized Medicine*, Vol. 9, No. 2, pp. 223-34.

The Daily Briefing (2011), "Will comparative effective research answer medicine's toughest questions?", 17 August, www.advisory.com/Daily-Breifing/2011/08/17/Will-comparative-effective-research-answer-medicines-toughest-questions, accessed 16 September 2013.

UK Department of Health (2012), "The power of information: Putting all of us in control of the health and care information we need", London, 21 May, www.gov.uk/government/uploads/system/uploads/attachment_data/file/213689/dh_134205.pdf, accessed 27 September 2013.

UN (2013), *World Population Prospects – The 2012 Revision*, United Nations, Department of Economics and Social Affairs, Population Division, DVD edition.

Vidal, M., M.E. Cusick and A.-L. Barabasi (2011), "Interactome networks and human disease", *Nature Reviews Genetics,* Vol. 12, pp. 56-68.

Wicks, P. et al. (2011), "Accelerated clinical discovery using self-reported patient data collected online and a patient matching algorithm", *Nature Biotechnology*, Vol. 29, No. 5, pp. 411-14.

Yin, J. et al. (2012), "ESA: Emergency situation awareness via microbloggers", paper presented at the Conference on Information and Knowledge Management (CIKM).

Further reading

Accenture (2012), "Patient access to electronic health records: What does the doctor order?", www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Patient-Access-to-Electronic-Health-Records-What-Does-the-Doctor-Order.pdf, accessed 2 August 2013.

AliveCor, www.alivecor.com/, accessed 8 October 2013.

Anwar, M.N. and M.P. Oakes (2011), "Data mining of audiology patient records: Factors influencing the choice of hearing aid type, BMC Medical Informatics and Decision Making", in *Proceedings of the ACM Fifth International Workshop on Data and Text Mining in Biomedical Informatics*, Association for Computing Machinery, pp. 11-18.

Asfaw, B. et al. (2010), "Host-based anomaly detection for pervasive medical systems", *Risks and Security of Internet and Systems (CRiSIS) 2010 Fifth International Conference*, pp. 1-8.

Astolfi, R., L. Lorenzoni and J. Oderkirk (2012), "Informing policy makers about future health spending: A comparative analysis of forecasting methods in OECD countries", *Health Policy,* Vol. 107, pp. 1-10.

Avron, J. (2007), "In defense of pharmacoepidemiology – Embracing the yin and yang of drug research", *New England Journal of Medicine,* Vol. 357, pp. 2219-21.

Barth, J. et al. (2012), "Combined analysis of sensor data from hand and gait motor function improves automatic recognition of Parkinson's disease", *Conference Proceedings*, Institute of Electrical and Electronics Engineers, Engineering in Medicine and Biology Society, pp. 5122-25.

Beniwal, S. and J. Arora (2012), "Classification and feature selection techniques in data mining", *International Journal of Engineering Research & Technology*, Vol. 1, No. 6.

British Medical Journal (2013), "Turning doctors into coders", Editorial, Vol. 347, 24-31 August.

Brothers, K.B., D.R. Morrison and E.W. Clayton (2011), "Two large-scale surveys on community attitudes toward an opt-out biobank", *American Journal of Medical Genetics Part A,* Vol. 155, No. 12, pp. 2982-90.

Burgel, P.P. et al (2010), "Clinical COPD phenotypes: A novel approach using principal component and cluster analyses", *The European Respiratory Journal*, Vol. 36, No. 3, pp. 531-39.

Cavouklan, A. (2013), "Looking forward: De-identification developments – New tools, new challenges", Office of the Information and Privacy Commissioner, Ontario, Canada, May.

Cecchini M. et al. (2010), "Tackling of unhealthy diets, physical inactivity, and obesity: health effects and cost-effectiveness", *The Lancet*, Vol. 376, No. 9754, pp. 1775-84.

Chaturvedi, S.K., V. Richariya and N. Tiwari (2012), "Anomaly detection in network using data mining techniques", *International Journal of Emerging Technology and Advanced Engineering*, Vol. 2, No. 5, pp. 349-353.

Chee, B.W., R. Berlin and B. Schatz (2011), "Predicting adverse drug events from personal health messages", *AMIA Annual Symposium Proceedings*, pp. 217-26.

Chellaprabha, B. and M. Archana (2013), "Anomaly data leakage detection", *International Journal of Engineering and Innovative Technology*, Vol. 2, No. 10, pp. 210-15.

Coloma, P.M. et al. (2012), "Electronic healthcare databases for active drug safety surveillance: Is there enough leverage?", *Pharmacoepidemiology and Drug Safety*, Vol. 21, pp. 611-21.

Computer Weekly (2013), "What does a petabyte look like?", www.computerweekly.com/feature/What-does-a-petabyte-look-like, accessed 17 September 2013.

Copeland, L.A., J.E. Zeber and C.P. Wang (2009), "Patterns of primary care and mortality among patients with schizophrenia or diabetes: A cluster analysis approach to the retrospective study of healthcare utilization", *BMC Health Services Research*, Vol. 9, No. 127.

Data Protection Working Party (2013), "Opinion 03/2013 on purpose limitation", adopted on 2 April, 00569/13/EN WP 203, http://idpc.gov.mt/dbfile.aspx/Opinion3_2013.pdf, accessed 23 April 2014.

Davenport, T.H. and D.J. Patil (2012), "Data scientist: The sexiest job of the 21st century", *Harvard Business Review*, October.

Department of Health and Human Services Centres for Medicare and Medicaid Services (2012), *Report to Congress: Fraud Prevention System First Implementation Year*, www.stopmedicarefraud.gov/fraud-rtc12142012.pdf, accessed 16 September 2013.

Di Iorio, C.T., F. Carinci and J. Oderkirk (2013), "Health research and systems' governance are at risk: Should the right to data protection override health?", *Journal of Medical Ethics,* www.ncbi.nlm.nih.gov/pubmed/24310171, accessed 15 March 2015.

Eddy, D.M. et al. (2011), "Individualized guidelines: The potential for increasing quality and reducing costs", *Medicine and Public Policy*, 3 May.

Eder, J., H. Gottweis and K. Zatloukal (2012), "IT solutions for privacy protection in biobanking", *Public Health Genomics,* Vol. 15, pp. 254-62.

El Emam (2013), *Risky Business: Sharing Health Data While Protecting Privacy*, Trafford Publishing, United States.

Enterprise Irregulars (2011), "The enterprise opportunity of big data: Closing the 'clue gap'", www.enterpriseirregulars.com/40616/the-enterprise-opportunity-of-big-data-closing-the-clue-gap/, accessed 17 September 2013.

EUBIROD (2013), "European best information through regional outcomes in diabetes", www.eubirod.eu/index.htm, accessed 20 September 2013.

EUnetHTA, European Network for Health Technology Assessment, www.eunethta.eu/, accessed 16 September 2013.

Eurobarometer (2010), *Biotechnology*, European Commission.

EUROCARE (2013), www.eurocare.it, accessed 20 September 2013.

European Commission (2003), Pharmaceutical Forum http://ec.europa.eu/enterprise/sectors/healthcare/competitiveness/pricing-reimbursement/european-initiatives/index_en.htm#h2-the-high-level-pharmaceutical-forum-on-pricing-and-reimbursement, accessed 16 September 2013.

European Commission (2012), Proposal for a regulation of the European Parliament and of the Council on the protection of the individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM(2012)0011 http://eurlex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=en&numdoc=52012PC0011, accessed 20 March 2013.

European Commission (2010), "Comparative study on different approaches to new privacy challenges, in particular in the light of technological developments", Working Paper No. 2: "Data protection laws in the EU: The difficulties in meeting the challenges posed by global social and technical developments", http://ec.europa.eu/justice/policies/privacy/docs/studies/new_privacy_challenges/final_report_working_paper_2_en.pdf, accessed 20 March 2013.

European Society of Cardiology, EURObservational Research Programme, www.escardio.org/eorp, accessed 16 September 2013.

Feldman, L. (2011), "Medical device integration – More than meets the eye", *For The Record*, Vol. 23, No. 13, 18 July, p. 20.

Foundation Centre, Number of Grantmaking Foundations 1975 to 2011, http://foundationcenter.org/findfunders/statistics/pdf/02_found_growth/2011/03_11.pdf, accessed 6 September 2013.

Global Fund to Fight Aids, "Tuberculosis and Malaria", www.theglobalfund.org/en/, accessed 6 September 2013.

Grand Challenges in Global Health, www.grandchallenges.org/Pages/Default.aspx, accessed 6 September 2013.

Health 2.0 Developer Challenge, www.health2con.com/devchallenge/blue-button-co-design-challenge/, accessed 20 September 2013.

Horgan, D., D. Byrne and A. Brand (2013), "EU directive on patients' rights to cross border healthcare: A largely theoretical achievement; so much more remains to be done, Editorials, *British Medical Journal*, Vol. 347, 7694f.

Hresko, A. and S.B. Haga (2012), "Insurance coverage policies for personalized medicine", *Journal of Personalized Medicine*, www.mdpi.com/2075-4426/2/4/201, accessed 8 October 2013.

Jaffe, M.G. et al. (2013), "Improved blood pressure control associated with a large-scale hypertension program, *Journal of the American Medical Association*, Vol. 310, No. 7, pp. 699-705.

Kaiser Permanente (2011), "Community benefit report: deeper and stronger, our commitment to total health continues to grow", http://share.kaiserpermanente.org/static/cb_annualreport/reports/docs/2011_CB_Annual%20_Report.pdf, accessed 13 September 2013.

Kimko, H.H.C. and C.C. Peck (eds.) (2011), "Clinical trial simulations: Applications and trends", *Advances in the Pharmaceutical Sciences Series*, American Association of Pharmaceutical Scientists, Vol. 1, Springer.

Lewis, C. et al (2013), "Public views on the donation and use of human biological samples in biomedical research: A mixed methods study", *British Medical Journal Open,* Vol. 3, No. 8, http://bmjopen.bmj.com/content/3/8/e003056.full, accessed 15 March 2015.

Lin, K.C. and C.L. Yeh (2012), "Use of data mining techniques to detect medical fraud in health insurance", *International Journal of Engineering and Technology Innovation*, Vol. 2, No. 2, pp. 126-37.

Liu, F., C. Weng and H. Yu (2012), "Natural language processing, electronic health records, and clinical research", in R.L. Richesson and J.E. Andrews (eds.), *Clinical Research Informatics (Health Informatics)*, Chapter 16, 293 Springer-Verlag London Limited.

Lusoli, W. et al. (2012) "Pan-European survey of practices, attitudes and policy preferences as regards personal identity data management", *JRC Scientific and Policy Reports*, European Commission.

Magoc, T. and D. Magoc (2011), "Neural network to identify individuals at health risk", *International Journal of Artificial Intelligence & Applications*, Vol. 2, No. 2, pp. 104-14.

Marx, V. (2013), "The big challenges of big data", *Nature*, Vol. 498, June.

Medcore Reveal Insertable Cardiac Monitors, www.medtronic.com/for-healthcare-professionals/products-therapies/cardiac-rhythm/cardiac-monitors-insert/reveal-dx-and-reveal-xt-insertable-cardiac-monitors-icms/index.htm, accessed 8 October 2013.

Microsoft Health Vault, www.healthvault.com/fr/en (accessed 8 October 2013).

Miksad, R.A. (2011), "When a decision must be made: Role of computer modelling in clinical cancer research*", Journal of Clinical Oncology,* 10 December, Vol. 29, No. 35, pp. 4602-4.

Mosen, D.M. et al. (2013), "More comprehensive discussion of CRC screening associated with higher screening", *American Journal of Managed Care*, Vol. 19, No. 4, pp. 265-71.

Mosen, D.M. et al. (2010), "Automated telephone calls improved completion of faecal occult blood testing", *Med Care*, Vol. 48, No. 7, pp. 604-10.

National Cancer Institute Fact Sheet, www.cancer.gov/cancertopics/factsheet/Risk/genetic-testing, accessed 8 October 2013.

National Centre for Health Statistics, National Health and Nutrition Examination Survey, www.cdc.gov/nchs/nhanes/response_rates_cps.htm, accessed 8 October 2013.

National Centre for Quality Assurance (2013), "NCQA to test pioneering way to measure quality, foster wider use of prevention strategies", press release, 9 April, www.ncqa.org/Newsroom/2013NewsArchives/NewsReleaseApril92013.aspx, accessed 27 September 2013.

National Institutes of Health, "Big Data to Knowledge Initiative", http://bd2k.nih.gov/about_bd2k.html#areas, accessed 23 September 2013.

NCBI Gene Expression Omnibus, www.ncbi.nlm.nih.gov/geo/, accessed 8 October 2013.

New York Times (2013), "Mining electronic health records for revealing health data", 14 January, www.nytimes.com/2013/01/15/health/mining-electronic-records-for-revealing-health-data.html?pagewanted=all, accessed 17 September 2013.

NIH Genetic Testing Registry, www.ncbi.nlm.nih.gov/gtr/, accessed 8 October 2013.

NIH National Human Genome Research Institute, www.genome.gov/10002335, accessed 8 October 2013.

OECD (2012), *Health at a Glance Europe*, www.oecd.org/els/health-systems/healthataglanceeurope.htm, accessed 7 May 2015.

OECD (2011), *Policy Issues for the Development and Use of Biomarkers in Health*, OECD Publishing, Paris.

OECD (2009a), OECD Guidelines on Human Biobanks and Genetic Research Databases, OECD Publishing, Paris.

OECD (2009b), *Pharmacogenetics: Opportunities and Challenges for Health Innovation*, OECD Publishing, Paris.

OECD (2007), *Genetic Testing: A Survey of Quality Assurance and Proficiency Standards*, OECD Publishing, Paris.

Open Government Partnership (2013), www.opengovpartnership.org/, accessed 27 September 2013.

Orphanet Report Series (2013), "Disease Registries" in *Europe*, www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf, accessed 18 October 2013.

O'Sullivan, B.P., D.M. Orenstein and C.E. Milla (2013), "Pricing for Orphan Drugs: Will the market bear what society cannot?", *Journal of the American Medical Association,* Vol. 310, No. 13, pp. 1343-44.

Patients Like Me, www.patientslikeme.com/, accessed 8 October 2013.

Pearson, J.F., C.A. Brownstein and J.S. Brownstein (2011), "Potential for electronic health records and online social networking to redefine medical research", *Clinical Chemistry* Vol. 57, No. 2, pp. 196-204.

Peptide Atlas, www.peptideatlas.org/, accessed 8 October 2013.

Rahm, A.K. et al. (2013), "Biobanking for research: A survey of patient population attitudes and understanding", *Journal of Community Genet*ics, October, Vol. 4, No. 4, pp. 445-50, http://dx.doi.org/10.1007/s12687-013-0146-0.

Rutter, C.M., D.L. Miglioretti and J. Savarino (2011), "Evaluating risk factor assumptions: A simulation-based approach", *BMC Medical Informatics Decision Making,* September, Vol. 7, No. 11, p. 55.

Simon, C.M., J.L. L'Heureux and B. Zimmerman (2011), "Active consent but not too active: Public perspectives on biobank consent models", *GENMED*, Vol. 13, No. 9, pp. 821-31.

Specimen Central (2013), www.specimencentral.com/biobank-directory.aspx, accessed 20 October 2013.

Statistics Canada, *Canadian Health Measures Survey Data Users Guide*, 2011.

Tatonetti N.P. et al (2011), "Detecting drug interactions from adverse-event reports: Interaction between paroxetine and pravastatin increases blood glucose levels", *Clinical Pharmacology & Therapeutics,* July, Vol. 90, No. 1, pp. 133-42, http://dx.doi.org/10.1038/clpt.2011.83 (accessed 15 March 2015).

Thompson Reuters (2013), Derwent World Patents Index, http://thomsonreuters.com/derwent-world-patents-index/, accessed 8 October 2013.

UCLA Centre for Networked Embedded Sensing, http://research.cens.ucla.edu/urbansensing/projects/find/, accessed 8 October 2013.

United Kingdom (2013a), Data.Gov.UK, http://data.gov.uk/, accessed 27 September 2013.

United Kingdom (2013b), www.gov.uk/government/news/health-secretary-to-strengthen-patient-privacy-on-confidential-data-use, press release, accessed 20 September 2013.

United States Environmental Protection Agency (2004), *Potential Implications of Genomics for Regulatory and Risk Assessment*, Washington, DC, December.

United States White House, Office of Science and Technology Policy, www.whitehouse.gov/blog/2013/06/07/health-datapalooza-iv-tops-huge-year-health-data-liberation-innovation, accessed 20 September 2013.

Ubaldi, B. (2013), "Open government data: Towards empirical analysis of open government data initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

Wang, C. et al. (2011), "Health and economic burden of the projected obesity trends in the US and the UK", *The Lancet,* Vol. 378, No. 9793, pp. 815-25.

Wang, X. et al. (2009), "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study", *Journal of the American Medical Informatics Association*, Vol. 16, No. 3, pp. 328-37.

World Medical Association (1964), *World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects*, archived at www.wma.net/en/30publications/10policies/b3/17c.pdf, accessed 7 May 2015.

Yeh, J.Y., T.S. Wu and C.W. Tsao (2010), "Using data mining techniques to predict hospitalization of hemodialysis patients", *Decision Support Systems*, Vol. 50, pp. 439-48.

Ylä-Herttuala, S. (2012), "Endgame: Glybera finally recommended for approval as the first gene therapy drug in the European Union, *Molecular Therapy*, Vol. 20, No. 10, pp. 1831-32.

Zhang, W. et al. (2011), "Role prediction using electronic medical record system audits", *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, pp. 858-67.

Zubi, Z.S. and R.A. Saad (2011), "Using some data mining techniques for early diagnosis of lung cancer", *Proceedings of the 10th WSEAS International Conference on Artificial intelligence, Knowledge Engineering and Data Bases*, pp. 32-37.

Zucchelli, E., A.M. Jones and N. Rice (2010), "The evaluation of health policies through microsimulation methods", *HEDG Working Paper*, University of York, January.

Chapter 9

Cities as hubs for data-driven innovation

This chapter provides an overview of data production and examples of opportunities for data-driven innovation in cities, as well as a discussion of related policy implications. The focus is on data-driven innovation i) that increases the efficiency of urban systems, including through system integration; ii) that enables new business opportunities, for example in urban mobility and accommodation markets; and iii) that improves urban governance. Examples in each of these areas show that the potential of data-driven innovation in cities has only begun to be tapped, and that the conditions to unleash it need to be improved. Issues to be addressed by policy makers to improve such conditions include interoperability, regulation, digital security risk management, and privacy.

> *"Cities have the capability of providing something for everybody, only because, and only when, they are created by everybody." (Jacobs, 1963)*

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

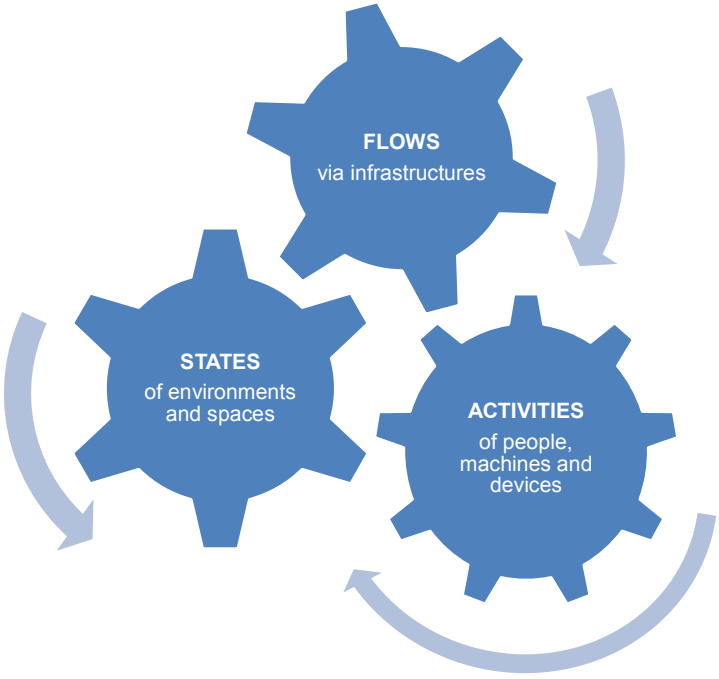
Sensors embedded in connected infrastructures, machines and devices, which are concentrated in urban areas, are producing an increasing variety and volume of data that are of significant worth for cities. A large share of the 65 million sensors estimated to be deployed worldwide in security, health care, the environment, transport and energy control systems today are embedded in urban infrastructures, facilities and environments (MGI, 2011). In US cities alone, an estimated 30 million CCTV cameras are installed in public spaces (Koonin, 2014). With around three-quarters of the OECD area population expected to be living in urban areas by 2022, cities will host at least 10 billion out of the 14 billion devices[1] estimated to be in use in member countries by then (OECD, 2010; OECD, 2012a).

9.1. The urban data ecosystem

Urban data production

The data produced in cities can be divided into three categories: data on flows, states and activities (Figure 9.1):

Figure 9.1. **Urban data categories**



- *Flows* – Cities are structured by and pervaded with different types of infrastructures (e.g. ICTs, transport, water, energy, waste networks) that facilitate movement and flows of resources, products, people and information across cities. Sensors embedded in urban infrastructures increasingly allow the digitisation and datafication of these flows. Some of the movement in cities is controlled by gateways, such as entrance doors, that are equipped with connected sensors, scanners, radio frequency identification (RFID) tags, etc.; these often require authentication to authorise entrance in specific areas.

- *States* – Urban inside spaces and outside environments are subject to constant natural and manmade changes. The particular state of urban spaces and environments – the density of people, air temperature and quality, light and sound levels, etc. – is increasingly monitored by sensors, including cameras; through synoptic instruments such as satellites; or in continual observations from urban vantage points. These states are being digitised and datafied in different forms and formats, including audio files, images, infrared and hyperspectral imaging, or radar.
- *Activities* – Connected machines and devices used for both personal and professional activities in cities allow measurement of transaction, consumption and communication patterns. These patterns include in particular: 1) people's activities, communication and interactions; 2) interaction between people and their environment; and 3) interactions among components of their environments, such as communicating and autonomous machines and devices. Furthermore, interactions and transactions of individuals and businesses with public institutions (e.g. tax records, land use, sales, inventories, public health, crime records, school outcomes, workforce development), with businesses (e.g. credit card payments, consumption behaviour, sales records), and individuals (e.g. social networking) create transactional data on activities in cities.

These data, created through the sensing, measuring and recording of flows, states and activities in cities, can also be distinguished by the extent to which they are location specific:

- data produced by stationary sensors embedded in urban infrastructures and environments, mostly describing flows and states in cities;
- geo-locational and geo-referenced data generated in cities, often from mobile devices and sensors, describing mainly the activities (actions, interactions, transactions) of connected people, machines and devices; and
- other data generated in cities that do not necessarily have geographic properties, such as data on financial transactions.

Actors in the urban data ecosystem

Many actors are involved in data collection and use in cities. Chapter 2 of this volume gives an overview of the data ecosystem as different layers and key actors (Figure 2.2): Internet service providers; IT infrastructure providers; data analytics providers; data providers; and data-driven entrepreneurs and innovators. While all of these actors are present in cities, this chapter mainly focuses on data providers and data-driven entrepreneurs and innovators.

The list below gives an overview of the key urban actors in the data ecosystem, which are the most relevant for the focus of this chapter. Each of them is in principle connected to all the others, through a digital layer and in multiple possible combinations. Each can be involved both in data collection and data use, at different times and in different functions:

- citizens and consumers
- innovators and entrepreneurs
- governments and utilities
- data brokers and platforms
- system operators and service providers.

The extent to which data can be exchanged among these actors and across systems in cities, as well as the extent to which they can easily be reused for different purposes, determines their potential for data-driven innovation (DDI) (see Chapter 2 of this volume). The main focus of this chapter is on data-driven innovations that i) increase the efficiency of urban systems, ii) enable new business opportunities, and iii) improve urban governance. By focussing on these topics, the chapter attempts to differentiate several issues that tend to be addressed in generalising discussions of the "smart city", with the aim to better understand the implications of DDI in cities for policy makers (Box 9.1).

Box 9.1. **"Smart City"**

The term "smart city" was mainstreamed by IBM's marketing campaigns and tends to cover a large range of different technologies, applications and services offered by companies like Siemens, Hitachi, General Electric, Cisco, Alstom, Arup, Microsoft and IBM itself, to name just a few of the biggest. The global smart cities market is estimated to grow to around USD 400 billion by 2020, technologies and services included (UKDBIS, 2013). A report by Navigant Research (2014) estimates that worldwide revenues from smart city technology will grow from USD 8.8 billion in 2014 to USD 27.5 billion in 2023.

An overwhelming number of cities have bought into the smart city narrative, much of which seems to remain a promise so far. In the EU-28, around 90% of cities with over 500 000 inhabitants identify themselves as a smart city. However, only half of these cities have actually implemented relevant initiatives, most of which are small in scale. That indicates the extent to which the smart cities market is characterised by a vendor push rather than by market and government pull (EP, 2014; Schaffers, 2011).

The term "smart" was defined in an earlier OECD report to be applicable when: "An application or service is able to learn from previous situations and to communicate the results of these situations to other devices and users" (OECD, 2013a). This definition can be applied to individual applications, services, machines and devices. However, it is unlikely to capture all aspects of the various systems that enable the functioning of cities, each of which is complex in itself and all of which potentially interact with each other. The term "smart" seems even less appropriate to capture the heterogeneity of a city beyond its technical components and systems, such as its human, social, environmental and economic realities.

Despite its widespread use, the term "smart city" does not provide a useful framework for analysing specific opportunities and policy implications of DDI in cities. It is thus not used in this chapter.

Source: UKDBIS, 2013; EP, 2014; Schaffers, 2011; OECD, 2013a.

9.2. Opportunities for data-driven innovation in cities

Efficiency gains

Much of the data on flows and states, and some of the data on activities in cities, can be used to increase the efficiency and promote integration of urban systems. The availability of historical and real-time[2] data on flows in transport, energy, water and waste systems enables analysis at unprecedented depth and granularity, as well as targeted interventions in and better management of urban systems. This section first looks at opportunities to improve the efficiency of urban systems; it then considers the potential for digitally integrating urban systems.

Opportunities in transport, electricity, water and waste systems

Promising effects of information and communication technologies (ICTs) and data use in cities can be found in transport. A main lever for data-driven improvements in transport is the direct match of demand and supply, based on fuller and often real-time information. There are, for example, mobile applications (apps), such as moovel or Citymapper, that inform urban travellers of the fastest connection from point A to point B, taking into account all available transport modes and traffic conditions. Matching demand and supply in real time allows shaving peak demand by redistributing it in space, reducing road congestion. This can save people time and money and reduce pollution and emissions in cities. Open data use in transportation, such as for apps providing real-time information on multimodal trips, prices and traffic conditions, is estimated to generate USD 720 billion to USD 920 billion per year (MGI, 2013).

Transport systems can be further optimised by dynamic road pricing and other types of traffic management based on real-time data analytics. Road pricing can be applied in different ways and for different objectives: dynamic road pricing to reduce peak time traffic (Singapore); dynamic parking fees to reduce the number of cars coming into congested areas (New York); or differentiated road pricing that favours environmentally friendly cars (Stockholm). Congestion charging in Stockholm has reduced traffic by 22% (100 000 passengers per day) and CO_2 emissions by 14% (25 000 tons annually) in the city centre, just in its seven-month trial period (CCLA, 2014; OECD, 2013b; KTH, 2010). The Intelligent Traffic Management System of London not only uses near-time traffic information to constantly adapt traffic light circuits, but also is able to learn from ongoing statistical observations of traffic patterns. It is becoming increasingly able to predict traffic and guide flows in anticipation of traffic volumes (TfL, 2010). The system is estimated to have reduced congestion in London by around 8% annually between 2014 and 2018 (TfL, 2011).

Electricity is another sector that can benefit from fuller capacity utilisation through data-based matching of demand and supply. Smart electricity grids are expected to yield energy savings for homes and businesses, in particular if combined with home and business energy management systems. Smart electricity meters inform households and businesses of real-time electricity prices based on current power demand and supply in electricity grids. Low prices incentivise consumption, whereas high prices discourage people to consume electricity. This levels out peaks in demand and thus reduces the necessary base load in the grid. Through the use of smart meters, European households are expected to save 10% of their energy consumption per year (EC, 2011). In the United States, the savings from smart grids are estimated to be 4.5 times the needed investment of USD 400 billion (EPRI, 2011).

DDI can furthermore yield efficiencies in water and waste systems. “Smart water solutions” are estimated to save water utilities USD 7.1 billion to USD 12.5 billion globally per year, through better i) leakage and pressure management techniques in water networks, ii) water quality monitoring, iii) smarter network operations and maintenance, and iv) data analytics in capital expenditure management (Sensus, 2012). Comprehensive and data-enabled approaches for waste reduction, recycling, reuse and waste-to-energy conversion can reduce energy consumption and emissions. For example, New York State’s “Beyond Waste” strategy estimated to save as much energy as consumed by 2.6 million homes each year (280 trillion BTUs) and to cut New York’s greenhouse gas emissions by around 20 million metric tons annually (DEC, 2014).

Potential synergies of urban system integration

Beyond improvements within separate urban systems, synergies can be reaped through integration of these systems. Understanding urban infrastructures and sectors as systems, a city can be considered a "system of systems", within which ICTs and the digitisation of urban flows are creating the potential for deep integration (CEPS, 2014). Already in an analogue world, urban systems were integrated to some extent; in Stockholm, for example, the transport, energy and waste systems are integrated to the extent that Stockholm's buses and taxis drive with biogas produced in the city's wastewater recycling plants (OECD, 2013b). Increasing digitisation of these systems will enable exchanging real-time information across the different systems involved, which in turn would allow optimising and scaling up such an approach. The same principle could be applied to other integrated systems, such as in the city of Kitakyushu, Japan, where industrial excess energy (heat) is reused to heat residential buildings through a district heating system that connects industrial with residential areas (OECD, 2013c). The use of real-time data on demand, supply and flows in urban systems can help deepen system integration and reap potential synergies of such integration.

A good example of a system that is becoming increasingly integrated with other urban systems, notably through the use of ICTs and real-time data, is the electricity grid. A key aspect of such "smart grids" is demand- and supply-side management, enabled by smart meters, that contributes to energy savings. A wider potential of smart grids, however, lies in integrating fluctuating renewable energy supply as well as electric vehicles. Electric vehicles can serve both for energy storage and supply, either at times of peak demand or in order to balance fluctuating renewable energy supply (OECD, 2012c; IEA, 2011). ICT-enhanced electricity systems thus enable a direct integration of transport and energy systems. Electricity grids can also be used to connect communicating devices in homes, and thus serve as an information system (OECD, 2013a). The "Internet of Things" will in any event multiply the systems, machines, devices and services connected via electricity grids and information systems, such as solar cells on roofs, weather stations, home heating systems and air conditioning, washing machines, light bulbs, electric vehicles, refrigerators, smartphones, supermarket stocks, etc. (see Chapter 2 of this volume).

Innovation hubs

Cities as living laboratories

The increasing collection and availability of data in cities have the potential to turn urban areas into large-scale experimental test beds for data-driven innovation. In contrast to many product and process innovations, large-scale system innovations – such as in transport or energy – require experimentation and testing at scale, ideally in real-life settings. Aiming to seize the opportunity of providing such settings, cities have started to define themselves as "living labs", such as the 340 European cities that are part of the European Network of Living Labs. This network defines urban living labs through four key elements: co-creation of new services by users and producers; exploration of emerging usages, behaviours and market opportunities; experimentation with implementing live scenarios within a community of lead users; and evaluation of concepts, products and services (Schaffers, 2011; ENoLL, 2014). Most urban living labs focus on providing the conditions for data-driven innovation, including through public-private and triple-helix collaborations (Box 9.2). The private sector has also discovered cities as ideal environments for DDI. Startupbootcamp accelerator programmes established in several European cities focus on DDI in mobile, near field communication, health and e-commerce; and companies like Microsoft have established incubators in cities like Paris, London and Berlin (Startupbootcamp, 2014; Microsoft Ventures, 2014).

Box 9.2. **Porto Living Lab and Guadalajara Creative Digital City**

Positioning itself as a living lab, Porto, Portugal aims to provide ideal conditions for DDI. At the core of the approach is an optical fibre backbone for high-speed Internet. The city collaborates with the University of Porto and an array of public and private stakeholders that constitute an institutional ecosystem for data-driven innovation. This ecosystem is open to researchers, private companies, public authorities and end users, which experiment and collaborate on data-driven products, services and applications, addressing specific challenges Porto is facing. Current projects include a platform for open data sharing, a machine-to-machine communication enhanced harbour management system, and a real-time traffic information service feeding connected buses and taxis in Porto.

The Ciudad Creativa Digital (CCD) Guadalajara, Mexico, is a joint effort of the Ministry of Economy, the governments of Jalisco and the City of Guadalajara, in coordination with the Massachusetts Institute of Technology (MIT) and private companies. Guadalajara aspires to set up digital infrastructures and an environment that attracts skilled creative human capital in order to develop digital content sectors and other innovative services. The city aspires to become a digital hub in Mexico and to serve as a living laboratory for DDI.

Source: Barros, 2013; Future Cities Project, 2014; Mexican Secretariat of Economy, 2015.

Beyond technical and institutional infrastructure, access to data is a key condition for fostering data-driven innovation in cities. In addition to giving access to city data via open data portals, many cities have started to directly incentivise data-driven innovation by rewarding programmers and entrepreneurs for developing data-driven applications. A common way of doing so is to organise hackathons, during which cities make data available to programmers, hackers and entrepreneurs and reward the most innovative applications, which usually are developed quickly. While these events tend to be very productive, so far they rarely have produced solutions that address deeper urban challenges. This seems to be partly due to a lack of focus in these events on actual challenges cities are facing (Townsend, 2014). Another shortcoming observed is a lack of resources to further develop promising applications and scale them up (Mulligan and Olsson, 2013). More recently, the private sector has started embracing the concept of hackathons and making private sector data publically available, such as during the Dutch Open Hackathon in Amsterdam (DutchOpenHackathon, 2014).

In recent years, many cities in OECD member countries have launched their open data portal. A City Open Data Census lists 70 US cities and provides metadata on their data sets (OKNF, 2014). In the European Union, over 120 open data initiatives and portals of cities or regions are listed, and a pan-European beta-version for a search portal (publicdata.eu) harvests metadata (EC, n.d.). In most cases cities publish structured (linked) data in machine readable formats to facilitate commercial and private use; however, few cities as yet offer application programming interfaces (APIs) (Open Cities, 2013). Many cities are using open source data portal platforms or software such as CKAN or Socrata, but no standards for open data portals exist so far.

New business opportunities

Over the past years, innovative start-ups have penetrated established urban markets with data-driven mobile apps and online platforms. Known under the label "sharing economy", these platforms allow people to rent ("share") cars, rides and bikes or space. They enable owners of assets and durable goods to turn them into services and thus make

excess capacity available for collective consumption. For example, car owners can rent their car if they are not using it, sell seats on trips they do anyway, or work as a private driver when time permits; real estate owners can rent living or commercial space whenever vacant. On the demand side, urbanites get more and cheaper mobility options, and travellers a larger and cheaper choice of accommodations; freelancers gain flexible access to office and commercial space. While this creates additional choices for consumers, it also raises questions with regards to quality control and insurance, and creates new competition for incumbents, notably for taxis and hotels.

People have shared cars and rides for a long time, but smartphones and real-time access to geo-locational data have allowed entrepreneurs to reinvent and scale up shared mobility commercially. Cars are among the most expensive and underused assets individuals own. On average, cars in cities are parked for 95% of their lifetime and are expensive: a US household spends USD 8 776 per year for its car, including gas, insurance, depreciation, vehicle payments and other expenses (ITF, 2012; *Time,* 2012). Car or ride sharing might not be cheaper per se, but they offer a flexible alternative to car ownership for many, in particular urbanites. And owners can top up their income by sharing their car or rides. The different variations of "sharing" in transport include private car rentals (Zipcar), ride sharing (Uber, Lyft, blablacar), and rentals of either free floating (Car2go, DriveNow) or station-based cars (Autolib'). Most of these services require subscription and are paid only if used. All transactions involved in using the service – from finding a ride or car to ordering and paying it – are taken care of by the online platform or mobile app. All participants in the service – drivers, car owners and passengers (renters) – can rate each other, which aims at creating trust, improving the quality of service and helps identify fraud or misuse. Similar principles are applied by other mobility apps, such as for shared parking spaces (justpark) and bicycles (Velib'). Studies on the potential effects of car and ride sharing in urban transport estimate that the size of car fleets in cities could be reduced significantly (Box 9.3).

Box 9.3. **Potential effects of shared mobility in urban transport**

Sharing cars, rides and bikes increases transport options in cities, can reduce resource consumption and has the potential to change the overall face of urban mobility. Ratti et al (2014) find that road mobility demand in Singapore could be met with only 30% of the vehicles currently in use in the city. A further 40% could be cut if all people on similar routes were to share their cars with others. Altogether, today's road mobility demand in Singapore could be met with about one-fifth of cars in the city (Ratti, 2014). A slightly more conservative calculation by the International Transport Forum estimates that car sharing could reduce the fleet size in cities by half and presents a scenario that combines high-capacity public transport with self-driving "TaxiBots" (self-driving shared vehicles) in which only 10% of cars would be needed (ITF, 2014).

These scenarios present a theoretically optimal version, which is not likely to be realised any time soon however. In the first place, shared mobility services could actually increase the number of cars on city roads, as early evaluations of car sharing systems have found. A main reason for this is that users of car-sharing services do not necessarily give up their private car, if they own one, and many users that sign up for car sharing offers did not own a car in the first place (Le Monde, 2013).

Given that it is early days for car and ride sharing systems, it is premature to judge their overall impacts on urban mobility. However, their successful adoption in many places and their economic potential indicate that their impacts will need to be considered: free-floating car-sharing systems alone are projected to generate annual revenues of EUR 1.4 billion in OECD cities with over 500 000 inhabitants by 2020 (Civity, 2014).

Source: ITF, 2012; ITF, 2014; Time, 2012; Ratti et al, 2014; Le Monde, 2013; Civity, 2014.

Another frontier in the "sharing economy" is the rental of living, working and commercial space via online platforms or mobile apps, mainly in cities. Again, home exchanges and temporary office rentals are nothing new, though the speed and scale at which short-term rentals of private spaces have become common practice is unprecedented. Similar to ride and car sharing, home sharing significantly builds on trust created by mutual ratings of landlords and guests as well as personal profiles and ID authentication. The online platform of Airbnb provides all basic services for the transactions between renters and guests, from advertising the place and securing direct communication between landlord and renter to providing a booking system, including payment, billing and insurance. Similar online platforms offer flexible office (ShareDesk) or shop rentals (Storefront), but are still small in scale. The former tend to be used by freelancers, the latter for pop-up stores**,** marketing campaigns or exhibitions. Home sharing in cities may affect local economies, however it is still unclear how (Box 9.4).

Box 9.4. **Potential economic effects of home sharing**

There is no comprehensive assessment yet of the economic effects of home sharing. However, anecdotal findings provide some insights. For example, for the case of New York, Airbnb claims that its guests are likely to generate more income for the city than hotel guests and that Airbnb guests tend to spend their money in areas which have traditionally not profited much from hotel guests and tourism.

The Airbnb study claims that in 2013, 416 000 visitors booked accommodation in New York through Airbnb, generating economic activity worth USD 632 million. On average, an Airbnb guest stayed 6.4 nights (compared to 3.9 for hotel guests) and spent USD 880 at NYC businesses (compared to USD 690 for average New York visitors). Most Airbnb listings in New York (82%) are outside the main tourist area of midtown Manhattan, compared to 30% of hotels located in these areas; and 57% of Airbnb visitors' spending occurs in the neighborhood in which they are staying.

While these figures give some indication about the behavior of Airbnb users, they do not represent a full picture of economic effects that Airbnb and other home sharing services have in a city. For example, there is no consideration of how home sharing affects the market share of hotels and the effects this could have on the local tax base and employment (Zervas et al., 2015). Neither is local spending of hotel employees taken into account, versus spending by Airbnb apartment owners, which are likely to be absent from the city, if they rent their entire home.

Source: Airbnb, 2014; Zervas et al, 2015.

Urban governance

Leveraging new sources of data

City administrations are increasingly using crowdsourced data to gain real-time and fine-grained information on public service delivery and infrastructure needs and conditions. Mobile apps like SeeClickFix in the United States or the BuitenBeter app in the Netherlands allow citizens to report on stray garbage, potholes, broken lamps and the like via their smartphone, directly to city hall. While this approach implies proactive citizens, mobile apps such as StreetBump in Boston report automatically; making use of the accelerometer (motion detector) and GPS, the StreetBump app reports on street conditions in Boston, notably on potholes and bumps, via drivers' smartphones. Also reporting automatically, the app Incidències 2.0 reports on commuter rail service

interruptions or delays in the metropolitan area of Barcelona; and the app Cycle Track informs transport planners in San Francisco of bicycle trips made throughout the city. The fine grained data collected through such mobile apps enables city governments to better target infrastructure investments, deliver tailored public services and increase efficiency in operations and maintenance. A more general account of possible uses of data by governments, including local governments, is given in Chapter 10 of this volume.

Online, crowdsourced and real-time data can also play an important role in disaster management in cities. For example, Crowdsense, a spin-off of the Dutch national applied research institute (TNO) and technical university (TU Delft), uses online and social media data for early warning systems and incident management services (Crowdsense, 2015). During the 2013 European floods, a citizen of Dresden, Germany, developed an online map that provided instant updates on flood hotspots and guided volunteers to places where help was most needed (DLI, 2013). In Japan, the Ministry of Internal Affairs and Communications (MIC) and the Tokushima Prefecture tested an evacuation system, which relies on IC cards, which are distributed to residents and linked to their home TVs. In case of an evacuation, the system displays the evacuation order with the residents' name on their TV screen. The residents' IC cards are scanned at the shelter to attain evacuees' up-to-date information. Furthermore, the system can recognise which residents watched TV just before an evacuation, so rescue workers can be dispatched to targeted houses. As a result, the evacuation time and the time to retrieve evacuees' information could be shortened significantly. While these examples give an idea of the multiple opportunities of using data for disaster management, such tools can only be one, albeit growing element within a city's risk and disaster management strategy, an increasingly important part of which in turn is digital security risk management, discussed below.

Applying data analytics

Crowdsourced, social media and other online data are increasingly used in city police departments, including for predictive data analytics and anticipatory decision making. CitiVox, a start-up that allows citizens to report crimes anonymously, provides governments with data from SMS and social media that can complement official crime statistics; policy makers and enforcement agencies can thus identify crime patterns they would not detect otherwise (CityVox, 2012). This is particularly useful in areas where fewer crimes are reported, such as in Central and South America. In the Netherlands, the application Buurt Bestuurt offers citizens the opportunity to engage with the city and other citizens in various ways to improve living conditions, including safety, in urban neighbourhoods (TNO, 2014). Some city police departments, such as in Los Angeles, Chicago, Memphis, Philadelphia and Rotterdam, are developing the capacity to analyse large and diverse data sets, including from social media, to support predictive policing. For example, data analytics may help identify potential future crime hotspots, prompting police to step up patrolling; or, it might be used to identify specific persons that are estimated to be prone to commit a future crime, as well as to determine surveillance levels for ex-prisoners. It should be noted that neither the effectiveness nor the privacy implications of such practices have as yet been thoroughly evaluated.

More comprehensive data on resource consumption allow policy makers to design more targeted and effective incentives to curb consumption – which may, however, have unintended consequences. For example, volumetric tariffs, such as applied for energy or water bills in many places, have proved successful in reducing resource consumption of households (OECD, 2012d). The increasing availability of data on other flows in cities allows similar models to be applied to other areas, such as waste disposal and recycling.

Recent research found that social network incentives could be a significantly more effective alternative to such traditional economic incentives. Instead of financially rewarding or punishing individuals for their actions (directly), as economic incentives do, social network incentives reward the friends of those who act. An experiment with incentives to save energy in Swiss municipalities has revealed that social network incentives were up to four times more effective than traditional incentive schemes (Pentland, 2014). While reducing resource consumption may be a desirable aim, the implications of nudging people's behaviours to become more rational are not yet well understood. As Frischmann (2014) points out, techno-social engineering might not only ignore the values of those being nudged by a modified "choice architecture", but also, if applied in institutional decisions, may generate path-dependency.

Greater data availability and more powerful computing are bringing urban modelling back to the forefront of urban planning. Urban modelling was first used over 50 years ago, but its imperfections – notably due to limited data and computing power – restricted its success. Its resurgence with the appearance of geographic information systems (GIS) in the 1990s and 2000s coincided with a shift from modelling aggregate equilibrium systems to complex, evolving "systems of systems" and urban dynamics (EUNOIA, 2012; Jin and Wegener, 2013). Geo-referenced data collected via crowdsourcing, remote sensing, social networking, smart transit ticketing, mobile phones and credit card transactions – combined with new computational power, including cloud computing – offer fresh possibilities for urban modelling, notably as applied to transport or integrated land use and transport planning (Nordregio, 2014). Opportunities for data-intensive urban modelling and simulations are being explored, both theoretically – as in the European EUNOIA project (Evolutive User-centric Networks for Intraurban Accessibility) – and practically. An example of the latter is the LakeSim project in Chicago, which has made extensive use of computational modelling to understand the impacts of alternative design, engineering and zoning solutions (UCCD, 2012). Data analysis and modelling of societal needs for urban infrastructures and services have the potential to significantly improve the pertinence of resource allocation and of investment decisions in urban areas.

9.3. Policy priorities

The extensive production and increasing variety and use of data in cities create great potential to spur DDI in urban systems, markets and governance. The extent to which such opportunities can be seized will depend considerably on policy makers at the national and sub-national levels. This section looks at some of the most important issues to be addressed.

Fostering interoperability

An important condition for integrating urban systems and advancing system-to-system communication is interoperability across different systems and components at different levels. International standards developed by standard-setting bodies – such as the International Electrotechnical Commission (IEC), the Institute of Electrical and Electronics Engineers (IEEE), the International Telecommunication Union (ITU) and the International Organization for Standardization (ISO) – will be crucial in scaling up the implementation of complex systems such as smart grids and in furthering integration of urban systems. Harmonised standards are key to achieving interoperability, at i) the *technical* level; ii) the *informational* (and semantic) level; and iii) the *organisational* level (CEPS, 2014). Many such standards do not as yet exist, but some have begun to be

developed. ISO currently works on a standard for smart community infrastructures, with the aim to foster complex system integration**.**

At the *technical* level, a major challenge comes with bringing a large number of companies, products and standards from different sectors into one increasingly integrated "system of systems". For example, the implementation of smart grids often necessitates large consortia of companies from domains that have not always collaborated before. The smart grid project in Issy-les-Moulineaux, Paris, for example, unites urban infrastructure actors such as Bouygues Immobilier, engineering and energy companies such as Alstom, EDF (Électricité de France), ERDF (Électricité Réseau Distribution France), Schneider Electric and Total, and communications and IT firms including Bouygues Telecom, Microsoft and Steria, in addition to various start-ups (CGDD, 2013). Beyond smart grids, important technical issues will need to be resolved for the joint functioning of mobile network architecture, ICTs, and Internet-based system architecture (Mulligan and Olsson, 2013).

Another technical challenge related to systems integration comes from the differing life cycles of different technologies, networks and infrastructures. This was pointed out in previous work by the OECD, which estimated life cycles (Table 9.1). Furthermore, many technologies are part of a complex legacy of existing networks and infrastructures, most of which were not designed for the data-driven applications or services discussed above.

Table 9.1. **Life cycles of selected technologies, networks and infrastructures**

Technologies, networks, infrastructures	Estimated life cycle in years
Consumer electronics	2-10
Home appliances	10
Vehicles	15
Telecom networks	10-50
Energy networks	15-50
Roads (maintenance)	30 (10)

Source: OECD, 2013a.

At the *informational* level, cross-sector data sharing is likely to pose challenges. Data collected in different sectors tend to be stored in different formats and few incentives exist for harmonising them. Without standards, data sharing may be limited by and locked into proprietary formats. Another issue for data sharing is privacy protection, which can be achieved to some extent through anonymising data before making it available. Some companies are proactive in this area. The company Orange, for example, uses the "Floating Mobile Data" technology to anonymise mobile phone data, which it offers for reuse, both for commercial purposes – such as in navigation systems and traffic management – and for research.

At the *organisational* level, the increasing need for data sharing will challenge vertical silos in public administrations and necessitates more co-operation among jurisdictions and levels of government. Co-ordination among different departments and agencies in public administrations has long been recognised as a crucial element of efficient and effective urban governance (Rodigo, Allio and Andres-Amo, 2009). These principles become even more pertinent in the context of cross-sectoral data sharing and data analytics (Koonin, 2014). Until now, few cities have provided good practice in this area. An exception is Barcelona, which has mechanisms in place to enhance co-ordination across different city departments as well as for public-private co-ordination on urban data and other horizontal issues. Furthermore, the multiple jurisdictions that make up large

urban areas, and multiple levels of government, need to be co-ordinated to improve data sharing. In the United States for example, city school districts, jails, criminal courts and public housing are often not under mayors' jurisdictions. Also, data from welfare programmes (e.g. Medicaid) are not usually available to cities. This means that US cities need to request specific agreements, which slows down or impedes the use of data and thus the potential for data-driven innovation and decision making in urban areas (Lane et al., 2014).

In the private sector, large companies tend to offer systemic and vertically integrated solutions, too much of which could lead to horizontal market separation. Some of the largest players in the "smart cities" market offer a systemic approach with proprietary solutions, ranging from sensor technology to the "city cockpit", in which all vital functions of a city should be monitored (Siemens, 2011). Other established firms in important sectors, notably in energy and communications, have started purchasing companies up- and downstream to gain control over larger parts of the value chain. Examples include AT&T Digital Life and the joint venture of Vodafone and British Gas with smart meters (OECD, 2013a). While vertical integration can to some extent help overcome vertical fragmentation in markets that are characterised by too many proprietary solutions along the value chain, it might lead to horizontal market separation and a lack of competition.

Reviewing legal and regulatory frameworks

New businesses labelled under the "sharing economy" have overcome high entry barriers and created new competition in established markets, notably in urban mobility and accommodation. Debates are ongoing in many places about how to react to new entrants like Airbnb competing with hotels, or Uber competing with taxis. In many places, these companies are operating in a legal and regulatory context which was shaped before their existence and that may need to be updated. Some countries and cities are trying to protect incumbents by punishing new entrants or making their activity illegal altogether. Others are reforming and providing clarity with new legislation and regulations. France, for example, passed a bill (ALUR) that allows renting both primary and secondary residences via Airbnb, and cities like Amsterdam or Hamburg have supportive policies towards Airbnb. Other cities, including New York and San Francisco, reacted with stricter regulation of the market. Regulators both at the national and sub-national level need to address the questions that arise through new business practices and consumer behavior, taking into account technological and societal developments that influence DDI.

Opening access to data

As highlighted in Chapter 4 of this volume, data can be considered as an infrastructure along with traditional infrastructures such as for transportation, communication or public utilities. The (re)use of such infrastructures typically generates positive spillover effects in particular when access is granted on equal and non-discriminatory terms (Frischmann, 2012). Recognising the value of public sector data for citizens, innovators and entrepreneurs, many cities have started to make their data available based on non-discriminatory access regimes such as through "open data". The interest in public sector data and its use to society are the subject of the OECD (2008) *Recommendation of the Council for enhanced access and more effective use of public sector information*, which suggests that governments implement the principles of openness, access and transparent conditions for reuse – and, where possible, for no or

marginal cost. City initiatives should be guided by these principles and be aware of the challenges related to opening data to the public.

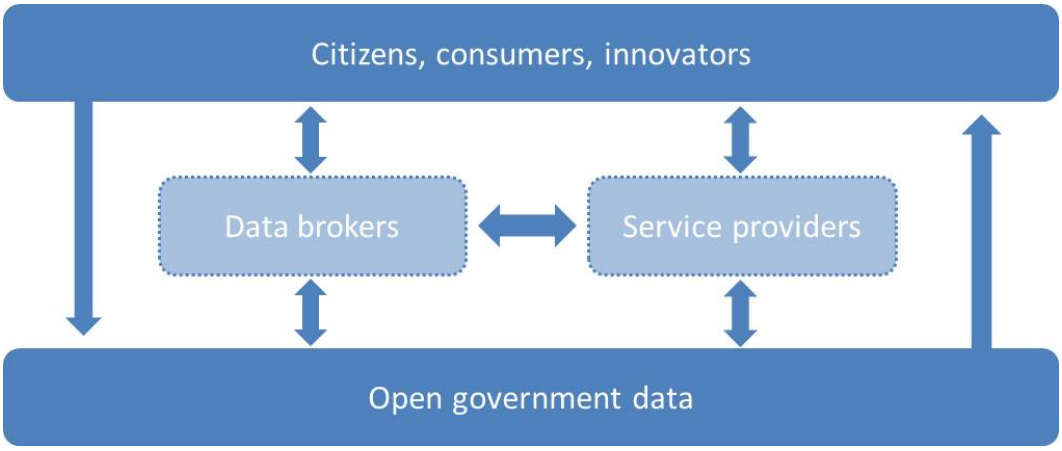
Opening access to data can be complicated. There are transaction costs stemming from agreements between different agencies; contractual and legal issues can arise from data collection; and existing rules are not adapted to data-driven service delivery or decision making in cities (Koonin, 2014). For example, in the Netherlands, Stadsbeheer, which uses citizen data collected via the BuitenBeter app to detect public incidents, cannot directly register this data in the Stadsbeheer back office, because that would infringe existing quality assurance protocols. Similarly, data from twitter and a more formal police app used in Rotterdam cannot be included in police reports for courts, given strict protocols the police have to follow.

Sensitive questions need to be addressed when it comes to what type of data cities should collect in the first place and what they should publish thereafter. Political considerations, regulatory frameworks, interests and values can influence the decision whether or not to collect and publish specific data (Kitchin, 2014). The University of Rotterdam, Netherlands, has developed a decision model to help urban policy makers determine whether and how a data set should be published for reuse (Gemeente Rotterdam, n.d.). In general, privacy guidelines[3] should be consulted; however, more specific discussions are needed – for example, on whether minimum requirements for city open data portals are useful, and which data should or should not be made public (Lane et al, 2014).

Another challenge for cities is to build the requisite capacity and skills for collecting, storing and analysing data in a depth and at a scale that are unprecedented, in addition to acquiring the infrastructure and computing power needed to store and process all the data. Best practice in the field of building capacity and skills are offered by the New York Mayor's Office of Data Analytics and that city's Center of Innovation through Data Intelligence or the Mayor's Office of New Urban Mechanics in Boston and Philadelphia, United States. Attracting data scientists to build in-house capacity will not be easy for many cities, notably given that similar skills are of great value in the private sector as well. With respect to infrastructure and computing power, many cities will not have the financial means or know-how to build and maintain local servers, and are likely to turn to cloud computing – which in turn raises new questions about security and privacy.

Private data providers, including data brokers, make additional data available but there is a lack of incentives and rules for providing private data to the public (see Chapter 4). Commercial data platforms like Esri or Sense-OS collect and (re)package data, make them publically available and provide analytical services. Focusing on data from the Internet of Things (IoT), the UK-based platform Thingful positions itself as a signpost for "the public IoT". Thingful helps find data and provides interaction among its users. Data brokers like Experian or Factual collect open and proprietary data and provide market intelligence; however, they do not necessarily facilitate interaction or data sharing among various interested players, mainly due to commercial interest (TNO, 2014). In other cases, regulation impedes private companies to make better use of their data in ways that could benefit the public. The interactions between open and propriety data are constantly evolving, and it might be too early to fix principles or rules to govern them. However, when starting to develop such principles, policy makers should be aware of the complex relationships among the actors engaged in producing, collecting, handling and using proprietary and open data (Figure 9.2).

Figure 9.2. **Key actors handling proprietary and open data in cities**



The multitude of actors involved in collecting and managing individuals' data in cities raises questions about the conditions under which data may be accessed and controlled by individuals. As a connected citizen, it has become difficult to know who is collecting, using, storing and sharing personal data where, when and with whom, and equally difficult to opt out of data collection. Only a few private companies that collect data from individuals enable data portability. Data portability would allow individuals to access their data when they end a contract with a firm at the latest, in order to keep it or use it in another context (Hemerly, 2013). Examples include start-ups like Handshake and Green Button: Handshake promises to keep individuals' data private, and allows them to hand those data out if they so wish for a reasonable price, and Green Button allows electricity consumers to have access to all data from their smart meters. Some public initiatives are aiming to offer services that move in this direction. For example, under the *Midata* initiative developed by the UK Government in co-operation with industry in the energy, finance, telecommunications and retail sectors, consumers will be provided with easier access to their consumption and transaction data in a portable and electronic format. This will enable them to gain insights into their own behaviour and make more informed choices about products that meet their interest. In France, Fing (Fondation Internet Nouvelle Génération) maintains MesInfos, an online platform through which consumers can access their financial, communication, health, insurance and energy data that are being held by businesses.

Better guidelines may be needed to ameliorate access to data throughout the economy and to help overcome existing barriers to data access, linkage and reuse. Existing frameworks that promote better access to data, some of which are sector specific, may need to be reviewed and eventually consolidated to foster coherence among public policies related, again, to data access, linkage and reuse. This would also include the OECD (2008) Council Recommendations promoting better access to data, including in particular the 2008 *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*, and the 2006 *Recommendation of the Council concerning Access to Research Data from Public Funding*, both of which are currently under review.

Managing digital security risks

Increasing digitisation of urban systems and digital system integration in cities can yield benefits, but it also creates new risks.[4] The more ICT, energy, transport and other critical urban infrastructures and systems are digitally interconnected, the more a city as a system-of-systems will become vulnerable to both internal and external threats, ranging from technical failures to cyber-attacks and natural disasters. For example, the flood caused by the 2012 Hurricane Sandy in New York City triggered a power blackout, which immediately affected critical urban infrastructures including transport and health, as well as telecommunications backhaul to over 2 000 cell sites in and around New York (Townsend, 2014). While in this case the shock that triggered system failures was a storm and the main system through which disruptions propagated into other systems was the electricity grid, in a different scenario the shock could come from a cyber-attack and disruptions would propagate through information and communication systems. Once the communication system, including the Internet, breaks down, an increasing number of critical urban functions will be affected. The fact that increasing digitisation and digital integration of urban systems will expose cities to new risks has been ignored by most cities so far (Cerrudo, 2015).

Critical urban infrastructures are becoming a key target for cyber-attacks. In 2013, the highest number of the US Industrial Control System Cyber Emergency Response Team's (ICS-CERT) responses in critical infrastructure sectors was in energy systems; sector specific on-site support by ICS-CERT (2011-13) concentrated on water and wastewater systems, transportation and energy (ICS-CERT, 2013). The majority of attacks address the digital component of the respective system. Israel Electric Corp. reported receiving around 6 000 attempted hacks per second on essential systems such as water, electricity, banking, rail and road infrastructures. In October 2012, for example, Haifa's traffic management system for a major artery in the city was hacked and caused hours of traffic chaos (Kitchin, 2014).

The increasing dependence of urban systems on digital functions and integration with other systems makes digital security risk management an important element for the economic and social development and resilience of cities. The core elements of a digital security risk management framework are addressed more in depth in Chapter 5 of this volume. Such a framework helps determine how to reduce risk to an acceptable level in light of the expected benefits, through security and preparedness measures that fully support the economic and social objectives at stake. Digital security risk management focuses on the uncertainties related to possible loss of the confidentiality, integrity or availability of digital activities that are becoming increasingly essential for the functioning of urban systems and services. In cities, digital security risk management should be fully integrated within overall risk management frameworks and approaches that address other types of uncertainties (i.e. not related to the digital environment). It should also take into account interdependencies among both digital and physical systems. The large number of actors involved will make co-ordination and co-operation across jurisdictions and levels of government – as well as among interdependent infrastructure, business and IT actors – crucial conditions for managing digital security risks.

Implementing privacy protection

Many of the opportunities and practices discussed in this chapter have implications for the protection of privacy. The framework for privacy protection in the context of data-driven innovation, elaborated in Chapter 5 of this volume, applies to cities as well. The

OECD (2013d) *Recommendation Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) should provide guidance to implementing the principles provided by this Recommendation, for sub-national policy makers much as for those at national or international levels. Policy makers need to step up efforts to implement privacy protection, in particular when the market provides insufficient incentives to protect personal data, and to find answers to questions that arise with new practices such as predictive analytics applied by city police departments. Using and linking large data sets, including personal data, to inform anticipatory decision making raises new privacy concerns that need to be addressed (White House, 2014). Finally, cities are likely to make increasing use of cloud computing, outsource storage and computing outside the city's jurisdiction and perhaps abroad. Potential issues of privacy protection in relation to cloud computing should therefore also be of concern to cities.

9.4. Key findings and policy conclusions

Cities are hubs for DDI, but this chapter has shown that the opportunities for DDI in cities have only begun to be tapped, and that policy makers play an important role in improving the conditions for DDI in cities. It can be noted in particular that:

1. *Urban systems can become more efficient through DDI, in particular through deeper system integration*. While separate urban systems such as for transport, energy, water and waste are already becoming more efficient through DDI, underexploited potential lies in deeper system integration. A fundamental condition for advancing such integration is interoperability among urban systems at technical, informational and organisational levels. Important enablers for interoperability are harmonised standards as well as multi-level governance and co-ordination across sectors and jurisdictions.
2. *Cities can be leveraged as laboratories for DDI*. Many cities have positioned themselves as "living laboratories" and new data-driven services are emerging, for example in mobility and accommodation markets. Two enabling conditions for DDI in cities are better access to urban data and appropriate review of legal and regulatory frameworks, taking into account technological and societal developments that influence DDI.
3. *DDI can improve urban governance*. Data collected from various sources, including crowdsourcing, can improve the evidence base for and precision of urban decision making. This applies to many areas, such as public service delivery, policing, incentive design, urban modelling, and disaster management. Two issues that need addressing by policy makers for each of these domains are digital security risk management and privacy; response to the latter could be guided by, inter alia, the principles of the OECD Privacy Guidelines.
4. *Better incentives and rules for sharing data*. Overall, the multiple public and private actors collecting and using data in cities need better incentives and rules for sharing that data in the interests of innovation. The design of such rules and incentives needs to balance both public and private interests, a challenge that policy makers have to face both at national and sub-national levels. This calls for coherence among open data frameworks, many of which relate to access, linkage and reuse (see Chapter 4 of this volume).

Notes

1 This does not take into account the increase in devices in non-OECD economies, nor the increase in devices for industrial applications.

2 In this chapter the notion "real time" stands in most cases for near-real time.

3 See the OECD (2013d) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines).

3 Risk is here understood as the effect of uncertainty on objectives.

References

Airbnb (2014), "Airbnb Annual Report", www.airbnb.com/annual/, accessed 19 September 2014.

Barros, J. (2013), "Cities in the cloud: A case study", presentation to the OECD, University of Porto and Veniam Works, 2 October.

CCLA (City Climate Leadership Award) (2013), "Award cities 2013: Barcelona, Amsterdam, Singapore", http://cityclimateleadershipawards.com/tag/intelligent-city-infrastructure/, accessed 13 May 2014.

CEPS (Centre for European Policy Studies) (2014), "Shaping the integrated infrastructures of cities", presentation in the IEC CEPS workshop "Orchestrating smart city efficiency", www.ceps.eu/system/files/article/2014/04/CEPS_IEC_Smart_City_3-Integrated_Infrastructures.pdf, accessed 04 November 2014.

Cerrudo (2015), **"**An Emerging US (and World) Threat: Cities Wide Open to Cyber Attacks", IOActive white paper, www.ioactive.com/pdfs/IOActive_HackingCitiesPaper_CesarCerrudo.pdf, accessed 23 April 2015.

CGDD (Commissariat Général au Développement Durable) (2013), "La ville intelligente: État des lieux et perspectives en France Développement", *Études et documents* Collection, Délégation au développement durable (DDD) of the CGDD, No. 73.

CityVox (2012), "Enhancing the government experience", CityVox presentation, http://siteresources.worldbank.org/INFORMATIONANDCOMMUNICATIONANDTECHNOLOGIES/Resources/D3S2bP4-OscarSalazar.pdf, accessed 22 March 2015.

Civity (2014), "Urban mobility in transition?", *matters No. 1*, Civity Management Consultants, Berlin.

Crowdsense (2015), "Early warning & incident management using social media", crowdsense website, http://crowdsense.co/, accessed 23 April 2015.

DEC (2014), "Climate smart waste management", Department of Environmental Conservation, New York State, www.dec.ny.gov/energy/57186.html, accessed 4 November 2014.

DLI (Deutschland Land der Ideen) (2013), "Hochwasserkarte Dresden", www.land-der-ideen.de/ausgezeichnete-orte/preistraeger/hochwasserkarte-dresden, accessed 19 September 2014.

DutchOpenHackathon (2014), "Dutch open Hackathon: Driving global innovation", www.dutchopenhackathon.com/en, accessed 19 September 2014.

EC (2011), "Next steps for smart grids: Europe's future electricity system will save money and energy", European Commission Press Summary,

http://ec.europa.eu/energy/gas_electricity/smartgrids/doc/20110412_press_summary.pdf, accessed 4 November 2014.

ENoLL (2014), About ENoLL, www.openlivinglabs.eu/aboutus, European Network of Living Labs, accessed 15 May 2014.

EP (2014), "Mapping smart cities in the EU", DG for Internal Policies, European Parliament, Policy Department A, Economic and Scientific Policy, EU, www.europarl.europa.eu/studies, accessed 19 September 2014.

EPRI (2011), "Estimating the costs and benefits of the smart grid", Electric Power Research Institute, http://ipu.msu.edu/programs/MIGrid2011/presentations/pdfs/, accessed 22 March 2015.

EUNOIA (2012), "Urban models for transportation and spatial planning", EUNOIA Consortium, www.nommon-files.es/working_papers/EUNOIA_PositionPaper_Oct2012.pdf, accessed 23 April 2015.

EC (n.d.), "Open data portals", European Commission, http://ec.europa.eu/digital-agenda/en/open-data-portals, accessed 19 September 2014.

Frischmann, B.M. (2014), "Human-focused turing tests: A framework for judging nudging and techno-social engineering of human beings", draft paper, 22 September, http://dx.doi.org/10.2139/ssrn.2499760, accessed 23 April 2015.

Frischmann, B.M. (2012), *Infrastructure: The Social Value of Shared Resources*, Oxford University Press.

Future Cities Project (2014), "Porto living lab: An ecosystem for the future", http://futurecities.up.pt/site/, accessed 19 September 2014.

Gemeente Rotterdam (n.d.), "Open data beslismodel", http://beslisboom.rotterdamopendata.nl/, accessed 19 September 2014.

Hemerly, J. (2013), "Public policy considerations for data-driven innovation", *IEEE Computer Society*, June.

ICS-CERT (Industry Control Systems Cyber Emergency Response Team) (2013), *ICS-CERT Year in Review*, National Cybersecurity and Communications Integration Center, Homeland Security, http://ics-cert.us-cert.gov/sites/default/files/documents/Year_In_Review_FY2013_Final.pdf, accessed 23 April 2015.

IEA (2011), *Impact of Smart Grid Technologies on Peak Load to 2050*, OECD/International Energy Agency, Paris.

ITF (2014), "Urban mobility: System upgrade", International Transport Forum and Corporate Partnership Board, http://internationaltransportforum.org/cpb/pdf/urban-mobility.pdf, accessed 23 April 2015.

ITF (2012), "Smart grids and electric vehicles: Made for each other?", Discussion Paper 2012-02, International Transport Forum, Paris, April.

Jacobs, J. (1963), *The Death and Life of Great American Cities*, Vintage Books, New York.

Jin, Y. and M. Wegener (2013), "Beyond equilibrium", *Environment and Planning B: Planning and Design*, Vol. 40, pp. 951-54. http://dx.doi.org/10.1068/b4006ge, accessed 23 April 2015.

Kitchin, R. (2014), "The real-time city? Big data and smart urbanism", *GeoJournal*, No. 79, pp. 1-14, http://dx.doi.org/10.1007/s10708-013-9516-8, accessed 23 April 2015.

Koonin, E.S. (2014), "The value of big data for urban science", in Lane, J., V. Stodden Bender S., and H. Nissenbaum (eds.) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press, United Kingdom.

KTH (Royal Institute of Technology) (2010), "Congestion charges which save lives", www.kth.se/en/forskning/sarskilda-forskningssatsningar/sra/trenop/trangselskatten-som-raddar-liv-1.51816, accessed 4 November 2014.

Lane, J., V. Stodden Bender S., and H. Nissenbaum (2014), *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press, United Kingdom.

Le Monde (2013), "On a raté l'objectif: Autolib' ne supprime pas de voitures", *Le Monde Blogs*, March, http://transports.blog.lemonde.fr/2013/03/26/on-a-rate-lobjectif-autolib-ne-supprime-pas-de-voitures/, accessed 19 September 2014.

MGI (McKinsey Global Institute) (2013), "Open data: Unlocking innovation and performance with liquid information", McKinsey & Company, www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information, accessed 22 March 2015.

MGI (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company, www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, accessed 22 March 2015.

Microsoft Ventures (2014), "Microsoft Ventures locations", www.microsoftventures.com/locations?r=4, accessed 19 September 2014.

Mulligan, C.E.A. and M. Olsson (2013), "Architectural implications of smart city business models: An evolutionary perspective", *IEEE Communications Magazine,* June.

Navigant Research (2014), "Smart technologies and infrastructure for energy, water, transportation, buildings, and government: Business drivers, city and supplier profiles, market analysis, and forecasts", Navigant Research, www.navigantresearch.com/research/smart-cities, accessed 19 September 2014.

Nordregio (2014), "Urban planning and big data – Taking LUTi models to the next level?", www.nordregio.se/en/Metameny/Nordregio-News/2014/Planning-Tools-for-Urban-Sustainability/Reflection/, accessed 19 September 2014.

OECD (2013a), "Building blocks for smart networks", *OECD Digital Economy Papers*, No. 215, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k4dkhvnzv35-en.

OECD (2013b), "Green Growth in Stockholm, Sweden", *OECD Green Growth Studies*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264195158-en.

OECD (2013c), "Green Growth in Kitakyushu, Japan", *OECD Green Growth Studies*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264195134-en.

OECD (2013d), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, 11 July, C(2013)79, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2012a), *Internet Economy Outlook*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264086463-en.

OECD (2012b), "Machine-to-machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

OECD (2012c), "ICT applications for the smart grid: Opportunities and policy implications", *OECD Digital Economy Papers*, No. 190, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9h2q8v9bln-en.

OECD (2012d), *OECD Territorial Reviews: The Chicago Tri-State Metropolitan Area, United States 2012*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264170315-en.

OECD (2010), *Cities and Climate Change*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264091375-en.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, C(2008)36, 30 April 2008, OECD Publishing, Paris, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2006), Recommendation of the Council concerning Access to Research Data from Public Funding, C(2006)184, 14 December 2006, OECD Publishing, Paris.

OKNF (2014), "US city open data census", Open Knowledge Foundation, http://us-city.census.okfn.org/, accessed 19 September 2014.

Open Cities (2013), "WP4 – Open Data", http://opencities.net/node/68, accessed 19 September 2014.

Pentland, A. (2014), *Social Physics: How Good Ideas Spread – The Lessons from a New Science*, Penguin Press, United Kingdom.

Ratti, C. and M. Claudel (2014), "The driverless city", www.project-syndicate.org/commentary/carlo-ratti-and-matthew-claudel-foresee-a-world-in-which-self-driving-cars-reconfigure-urban-life, accessed 15 May 2014.

Rodigo, D., L. Allio and P. Andres-Amo (2009), "Multi-level regulatory governance: Policies, institutions and tools for regulatory quality and policy coherence", *OECD Working Papers on Public Governance*, No. 13, OECD Publishing, Paris, http://dx.doi.org/10.1787/224074617147.

Schaffers, H. et al. (2011), "Smart cities and the future Internet: Towards cooperation frameworks for open innovation", in G. Goos et al. (eds.), *The Future of the Internet – Future Internet Assembly 2011: Achievements and Technological Promises*, Springer, pp. 431-47.

Sensus (2012), "Water 20/20", Sensus, http://sensus.com/documents/10157/1577608/Sensus_Water2020-USweb.pdf/d67d0a75-255a-4a20-86f1-d4548bfcdf78, accessed 4 November 2014.

Siemens (2011), "Real-time government", www.siemens.com/innovation/apps/pof_microsite/_pof-spring-2011/_html_en/city-cockpit.html, accessed 19 September 2014.

Startupbootcamp (2014), "Startupbootcamp Accelerator Programs, www.startupbootcamp.org/accelerator.html, accessed 19 September 2014.

TfL (Transport for London) (2011), "London's intelligent traffic system", presentation by the Director of Strategy, Surface Transport for London, www.impacts.org/euroconference/barcelona2011/Presentations/11_Keith_Gardner_presentation_Barcelona_v2.pdf, accessed 23 April 2015.

TfL (2010), "Traffic control and signal modelling in the capital: A 2020 approach in 2010", presentation at the Roads Summit, www.scoot-utc.com/documents/SA_Roads_Summitv2.pdf, accessed 23 April 2015.

Time (2012), "What's car sharing really like?", *Time Business*, April, http://business.time.com/2012/04/16/whats-car-sharing-really-like/, accessed 19 September 2014.

TNO (2014), Data and the City, *TNO report*, http://publications.tno.nl/publication/34610049/PkRiC6/TNO-2014-R10951.pdf, accessed 13 May 2015.

Townsend, A.M. (2014), *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*, W.W. Norton & Company, New York and London.

UCCD (2012), "LakeSim: A prototype workflow framework for coupling urban design and computational modeling tools", Urban Center for Computation and Data, https://urbanccd.org/projects/lakesim-prototype-workflow-framework-coupling-urban-design-and-computational-modeling-tools, accessed 19 September 2014.

UKDBIS (2013), "The smart city market: Opportunities for the UK", Bis Research Paper No. 136, UK Department for Business Innovation & Skills.

White House, United States (2014), "Big data: Seizing opportunities, preserving values", Executive Office of the President, May, www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, accessed 23 April 2015.

Chapter 10

Governments leading by example with public sector data

This chapter examines the benefits and challenges of opening access to data from one of the economy's most data-intensive sectors, the public sector. The potential of public sector information (PSI) including open government data (OGD) is discussed from several perspectives: use by government itself which, in tandem with data analytics, can make for better informed policy making and enable delivery of more innovative services; open access for citizens, which can greatly improve accountability through transparency and lead to citizens' empowerment; and reuse in the private sector, a stimulus to innovation. The challenges in implementing open data strategies are also enumerated, including dissuasive pricing and licensing practices; differences in licensing systems across national institutions; lack of information and standards and poor interoperability; organisational and cultural obstacles within the public sector; and legal constraints. The chapter concludes with a number of recommended policy options.

A leader is best when people barely know that he exists, not so good when people obey and acclaim him, worst when they despise him. Fail to honor people, They fail to honor you. But of a good leader, who talks little, when his work is done, his aims fulfilled, they will all say, "We did this ourselves." (Lao-Tzu, Tao Te Ching)

« L'ouverture et le partage des données publiques ne sont pas vus comme des fins en soi, mais comme des leviers qui peuvent être mis au service de trois objectifs : une démocratie plus aboutie ; l'innovation et la croissance ; et une meilleure efficacité de l'action publique. » (Verdier, 2014)

Public organisations produce and collect a huge volume of data in order to perform their tasks, making the public sector one of the economy's most data-intensive sectors (OECD, 2013). In the United States, for example, public sector agencies stored on average 1.3 petabytes (millions of gigabytes) of data in 2011, and the public sector is that country's fifth most data-intensive sector. Chapter 2 has highlighted the public sector as an important actor in the data ecosystem, in both respects: as a key user of data and analytics, and as a key producer of data that can be reused for new or enhanced products and processes across the economy – that is to say, for data-driven innovation or DDI.

Better access to and use of public sector data can lead to important value creation from economic, social, and good governance perspectives (Vickery, 2012; Ubaldi, 2013; OECD, 2015). Direct use of public sector data can generate products and services, and thus contribute in a variety of ways to improved efficiency and productivity within the public sector and across the economy. Public sector data can thus contribute to the shift towards knowledge-based societies and economies, where data is a potential driver of growth, employment, as well as of improved public service delivery and more efficient, transparent and participatory governance.

The economic value here is certainly significant: the value of the OECD market for public sector information – PSI, including public sector data – was estimated to be around USD 97 billion in 2008, and could have grown to around USD 111 billion by 2010. Aggregate OECD economic impacts of PSI-related applications and use were estimated to be around USD 500 billion in 2008, and there could be close to USD 200 billion of additional gains if barriers to use are removed, skills enhanced, and the data infrastructure improved. These are among the Principles reviewed in the OECD (2008) *Council Recommendation for Enhanced Access and More Effective Use of Public Sector Information* (OECD PSI Recommendation; see Annex and OECD, 2015 on the review of its implementation by governments).

In 2013 the OECD conducted a survey focusing on open government data, or OGD (see Box 10.1 defining PSI and OGD).[1] The intention was to acquire a comprehensive picture of national efforts and contexts for open data implementation. The knowledge base created could then serve as an indicator of countries' progress in developing metrics on OGD impact and value creation. Aspects covered by the survey were strategic approach; implementation efforts; countries' focus on value and impact creation; and the main challenges for further progresses. The survey reveals many of those data sets most generally available in OGD portals are commercially valuable. These include (in decreasing order of citation): meteorological and environmental information (19 out of 20 countries cited these as available), geographical information, social information, cultural information and content (each cited by 18 out of 20 countries), economic and business information, traffic and transport information, tourist and leisure information, and educational information. Countries with high data set availability by domain are: Canada, Denmark and France (all 15 domains listed including selected defence areas), Australia, New Zealand, Slovenia (all domains except defence) and the United Kingdom. Countries with the lowest number of domains available are the Netherlands and Portugal (7 out of the 15 listed domains), Italy (9 out of 15), Germany and Norway (Figure 10.2).

Box 10.1. **Defining public sector information and open government data**

Public sector information (PSI) is information (including data) generated by the public sector as part of its public task; the term covers weather, map, statistical and legal data, as well as digital content held and maintained by the public sector in galleries, libraries, archives and museums. PSI is increasingly made available through open access regimes, as specified by the Openness Principle of the *OECD (2008) Council Recommendation for Enhanced Access and More Effective Use of Public Sector Information* (PSI), often at low or no cost. PSI is made available for potential reuse for economic and social ends that for the most part are not within the public sector or aimed at enhancing government services. Nevertheless, government efficiency and effectiveness is improved by easier information access and transfer across agencies at low or no cost and without restrictive legislative controls.

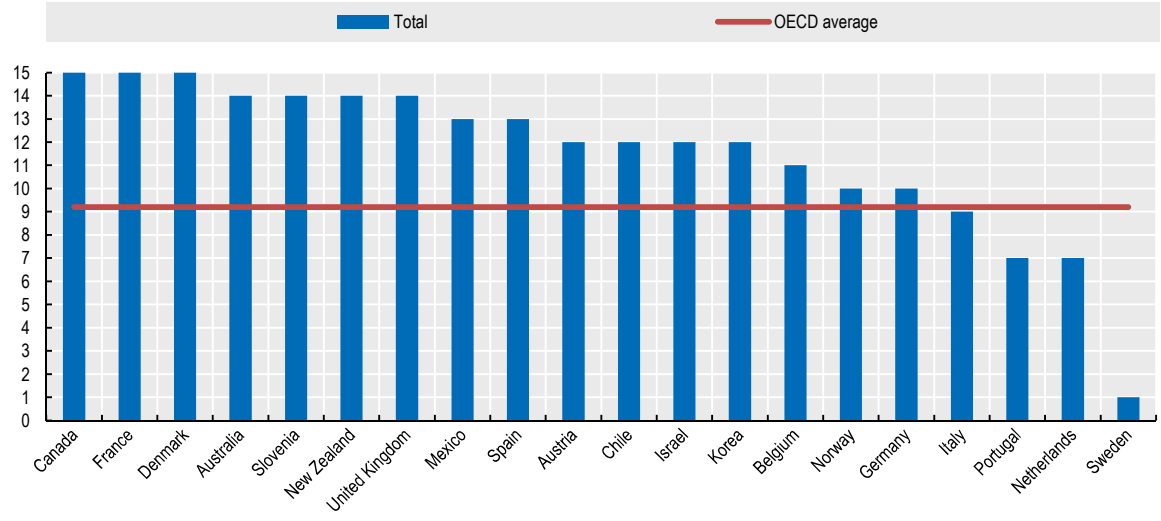
Open government data (OGD) refers to government or public sector data (i.e. any "raw" data produced or commissioned by the public sector) made available through open access regimes, so that they can be freely used, reused and distributed by anyone, subject only to (at the most) the requirement that users attribute the data and (sometimes) that they make their work available to be shared as well. Enhancing transparency, accountability and citizen participation for good governance and socio-economic development is an important objective of OGD (OECD, 2014a). Open government data are a subset of PSI (Figure 10.1); PSI, in addition, includes not only data but also digital content, such as (e.g.) text documents and multimedia files. In this chapter, the terms "open government data" and "public sector data" are synonymous.

Figure 10.1. **The relationship between public sector information and open government data**



The key idea behind open access to public sector data is that value can be derived through the reuse of that data by any user from within or outside the public sector. As highlighted in Chapter 4, the full range of goods and services enabled by data often cannot be anticipated ex ante by the data producer, and thus may not be realised if access is limited. Given that data are a non-rivalrous good, social welfare is maximised when everyone who values the data can use them. Therefore, the gains from public sector data emerge from removing any type of disincentive to data access and reuse.

Figure 10.2. **Variety of data sets in the centralised government portal**



Note: Data refer to the number of different types of data provided in the centralised portal in a list of 15 policy domains: economic and business information, geographic information, legal system information, meteorological and environmental information, social information, traffic and transport information, tourist and leisure information, agricultural, farming, forestry and fisheries information, natural resource information, scientific information and research data, educational information and content, public order and safety information, defence (including military), political information and content, cultural information and content.

Source: OECD survey on Open Government Data, version 1.0, 19 April 2013.

Open access to public sector data comes with great promise but also with many barriers and challenges. As discussed below, impediments include dissuasive pricing and licensing practices; differences in licensing systems across national institutions; lack of information and standards and poor interoperability; organisational and cultural obstacles within the public sector; and legal constraints impeding easy access, use, reuse and data sharing within and across levels of government. Additionally, as public sector data are progressively seen as potential public value generators for society (including the public and private sector) – rather than as a source of government revenue – pricing practices are moving towards making data accessible for free or at a marginal cost (e.g. the cost of reproducing the data when necessary). However, at times of budget cuts and financial constraints, governments feel the need to clearly articulate a business case and identify funding models to open up and digitise government data without penalising data providers. For similar reasons, great emphasis is now placed on devising more solid methodologies to assess the impact of open access to public sector data.

The purpose of this chapter is to discuss the potential of public sector data from a public sector perspective, highlighting some of the main trends in open government data and PSI strategies and initiatives in OECD member countries. The chapter first highlights the potential of DDI, focusing on the potential of public sector data analytics for the public and private sector and for citizens. It then discusses key challenges for implementing open data in the public sector. Analysis of government strategies for implementation in several OECD countries follows, and the chapter concludes with key findings and policy conclusions.

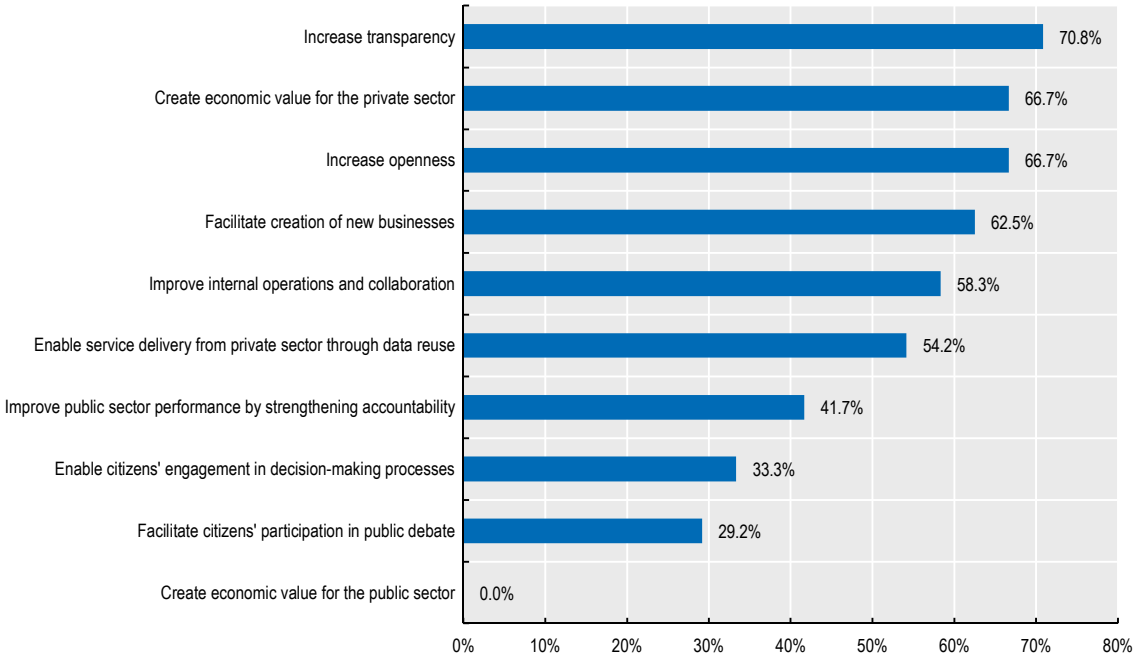
10.1. The potential of public sector data

As a recent OECD survey highlights, there are a number of objectives linked to enabling the reuse of data. These include broader societal and economic aims than simply

producing value internally within the public sector (Figure 10.3). In particular, creating economic value for the private sector ranked among the top objectives noted in the survey, no matter how they are counted or regrouped. Objectives related to good governance value – i.e. transparency and openness – also rank at the top. This is consistent with the fact that many national agendas on open data emanated as complement, or reinforcement, of national transparency agendas. The objectives of citizen participation and citizen engagement ranked lower than would be expected, given that many open government and service delivery agendas point to open data as a key enabler for strengthened public engagement in service design, policy making and rulemaking.

Figure 10.3. **Main objectives of open government data strategies**

Percentage of countries ranking each feature among their top five objectives



Source: OECD survey on Open Government Data, version 1.0, 19 April 2013.

This section looks at the benefits and value that can be derived from the use of public sector data – for governments (e.g. public sector productivity and internal costs savings, improved policy development, more effective service delivery, transparency), for citizens (e.g. public participation and engagement, people's empowerment), and for businesses (e.g. product and process innovation).

Use of public sector data and analytics by government

Due to the recession and government budgetary constraints, governments feel an urgent need to improve their own performance. This involves making the transfer of information and data among different parts of government more efficient, transparent and less costly; reducing or eliminating the burden of inter-agency charging for data; and developing common data access, all to achieve public sector productivity gains and more effective service delivery.

The use of OGD by government agencies can lead to efficiency improvements in the public sector. It can, for example, help bring down silos and foster collaboration across

and within public agencies and departments. As common or shared data sets and/or registers are being created, collaboration and exchange on who owns which public information and for what purpose are needed. This provides an opportunity to re-engineer and simplify internal procedures and/or delivery services in new ways. Moreover, public resources are freed from having to maintain individual registers, and data sets can be reallocated to more productive tasks. Finally, the release of public sector data has provided a platform for innovation in service delivery, as will be seen below. This has resulted from the reuse of data not only by private sector actors or civil society organisations, but also by civil servants who have in several instances taken the initiative – for example, to develop new apps.

Furthermore, the increasing amount of data made available in formats that enable reuse and linkage is supporting the expansion of data analytics in the public sector. As is the case with the private sector, the use of data and analytics holds great potential for value creation for the public sector. Predictive data analytics can, for example, facilitate identification of emerging governmental and societal needs. And greater ability to combine different (public and private) data sets can help develop enhanced insights that can be used for innovative goods and services. Authorities point to the need in the future not just for "big data" to draw on citizen data and facilitate analytics – for example to develop and simulate public policies and better target public services – but also for a more qualitative approach that includes ethnographic surveys. Use of this data by the public sector can also make for better decisions, inform policies, support the development of data-driven processes and services, and deliver more innovative services (Ubaldi, 2013). There are two major types of DDI benefits for the public sector: i) improvement of the evidence base for policy making, and ii) personalised public service delivery.

There are also, of course, considerable risks in governments' use of data analytics, in particular with regard to the privacy of citizens. Advances in analytics make it possible to infer sensitive personal information that citizens may not even have shared with governments. This is especially the case when data from different sources are linked across public sector bodies, or with personal data available on the Internet. As highlighted in Chapter 5, misuse of these insights can affect core values and principles, such as individual autonomy, equality and free speech, and may have a broader impact on the functioning of democratic societies as a whole. For example, while personalisation enabled by data analytics may result in greater efficiencies for public service delivery as highlighted below, it may also lead to discrimination that limits citizens' ability to escape the impact of pre-existing socio-economic indicators. Governments should therefore lead by example by seriously addressing the privacy challenges when using data analytics for the benefits presented below. Possible responses discussed in Chapter 5 include improving transparency, better access and empowerment of citizens, promoting responsible usage of personal data by organisations, and the use of technologies in the service of privacy protection. Finally, application of risk management to privacy protection may effectively protect privacy through ongoing DDI, including in the public sector.

Personalised public service delivery

Over the past decade the private sector has increasingly used data analytics to target the delivery of goods and services. There is much governments can learn from the private sector regarding methods for combining the use of (personal) data and the latest technology to achieve targeted delivery. True citizen-government dialogue, however, requires structuring, tracking, tracing and personalising answers to the input received by local officials, and at the right level in the government rather than by an anonymous agency or ministry. This

demands time and effort, but can also mean wins for citizens as well as for government. It can move governments from one-size-fits-all to segmentation and finally to personalisation. Estimates suggest that better exploitation of data could significantly increase efficiency, with billions of EUR in savings for the public sector. According to MGI (2011), full use of big data in Europe's 23 largest governments may reduce administrative costs by 15-20%, creating the equivalent of EUR 150 billion to EUR 300 billion in new value, and accelerating annual productivity growth by 0.5 percentage points over the next ten years.[2] The main benefits are estimated to come from greater operational efficiency (due to greater transparency), increased tax collection (due to customised services, for example), and fewer frauds and errors (due to automated data analytics) (see Box 10.2).

Box 10.2. **Data analytics at the municipal level**

In New York City (NYC), data analytics promises to better target fire, safety and health inspections. NYC receives over 20 000 complaints per year for "illegal conversion", i.e. properties that house more people than is considered safe. Historically, inspectors at the Department of Buildings (numbering around 200) would find serious high-risk conditions at 13% of inspections. Recently, the Department embarked on co-operation with around 20 other NYC agencies. They cross-tabulated enormous amounts of additional data on the individual properties, and used the results to guide inspections. The result is that currently, between 70% and 80% of inspections discover high-risk properties, for which action can be taken. Moreover, the NYC mayor office used advanced analytics and combined data from several of the city's departments to boost predictive capacity and help save lives and taxpayers' money in the city. Results include:

- a fivefold return on the time building inspectors spend looking for illegal apartments
- an increase in the rate of detection of dangerous buildings that are highly likely to result in firefighter and tenant injury or death
- more than a doubling of the hit rate for discovering stores selling bootlegged cigarettes
- a fivefold increase in the detection of business licences being flipped
- fighting the prescription drug epidemic through detection of the 21 pharmacies (out of an estimated total of 2 150 in NYC) that accounted for more than 60% of total illegal Medicaid reimbursements for oxycodone in the city.

Similar studies focusing on the United Kingdom show that the public sector could save GBP 2 billion in fraud detection and generate GBP 4 billion through better performance management by using big data analytics (Cebr, 2012). That does not include the potential for public health care, where for example analytics is used for diabetes audit data.[3] The National Diabetes Audit toolkit analyses data originating from primary care that are linked with secondary care data sources. Data can be stratified and analysed in many different ways, e.g. sex, age, ethnicity. The information comes from general practitioner (GP) practices, primary care trusts (PCTs), strategic health authorities (SHAs) and hospital diabetes units, specialist paediatric units and HES/PEDW (Hospital Episode Statistics/Patient Episode Database for Wales). The potential of data analytics in the health care system is discussed in detail in Chapter 8.

Combining public sector data and external data sources for policy making

Torrents of data streaming across public and private networks can improve the quality of statistics in an era of declining responses to national surveys, and can create close to real-time evidence for policy making in areas such as prices, employment, economic output and development, and demographics (Reimsbach-Kounatze, 2015). Not just for OECD countries but also for developing economies, the exploitation of these new data sources (through public-private co-operation) provides a new opportunity to better inform public policy making (UN Globalpulse, 2012).[4]

Among the new sources of statistics that policy makers are now using as a complement to existing public sector data are search engine data derived from keywords entered by users searching for web content. Google Insights for Search, for example, provides statistics on the regional and time-based popularity of specific keywords (see Chapter 3). Where keywords are related to specific policy topics such as unemployment, Google Insights can provide real-time indicators for measuring and predicting unemployment trends that policy makers are increasingly considering as a complementary statistical source.[5] The Central Bank of Chile, for example has explored the use of Google Insight for Search to predict present (or "nowcast") economic metrics related to retail good consumption (Carrière-Swallow and Labbé, 2010).

Other statistics are created by directly "scraping" the web. The Billion Price Project (BPP), for example, collects price information over the Internet to compute a daily online price index and estimate annual and monthly inflation. The index is basically an average of all individual price changes across all retailers and categories of goods. More than half a million prices on goods (not services) are collected every day by "scraping" the content of online retailers' websites such as Amazon.com. This is not only five times what the US Government collects, it is also cheaper because the information is not collected by researchers who visit thousands of shops, as they do for traditional inflation statistics. Also, unlike official inflation numbers that are published monthly with a time lag of weeks, the online price index is updated daily with a lag of just three days. In addition, the BPP has a periodicity of days as opposed to months. This allows researchers and policy makers to identify major inflation trends before they appear in official statistics. For example, in September 2008, when Lehman Brothers collapsed, the online price index showed a decline in prices, a movement that was not picked up until November by the consumer price index (Surowiecki, 2011). Governments in the United States, the United Kingdom, Germany and France, and in key Partner countries such as Brazil, have established a partnership with PriceStats, which manages the BPP index, to contribute to and use the index.

Rapid take-up of these new sources by policy makers is a growing trend, although it should be acknowledged that methods to mine the sources are still in their infancy and need rigorous scientific scrutiny. Besides the privacy challenges highlighted above, there are considerable risks that the underlying data and analytic algorithms could lead to unexpected false results – an even greater danger when decision-making is automated (see Chapter 3). Governments should therefore be aware of the limitations that come with the use of data and analytics; their activities could otherwise be based on wrong assumptions and lead to social and economic harms to citizens. A number of national statistical offices (NSOs) are currently exploring, if not already tackling, the benefits and challenges of supplementing official statistics with big data. In September 2013, for example, the European Statistical System Committee (ESSC) adopted the Scheveningen Memorandum on Big Data and Official Statistics (ESSC, 2013) to encourage partners of the ESSC to "effectively examine the potential of Big Data sources" and to "adopt an

action plan and a roadmap by mid-2014". As another example, the High-Level Group for the Modernisation of Statistical Production and Services (HLG), which was set up by the Bureau of the Conference of European Statisticians to promote standards-based modernisation in 2010, began to assess the potential of "big data" in 2014.[6]

Improving government accountability, transparency and responsiveness as well as democratic control

Strong supporters of open data as a key enabler of open government believe there is a correlation between lack of open government data and levels of corruption in any given country. For instance, a common assumption is that the lack of data in the public domain allows public servants to engage in corrupt behaviour with impunity. In addition, open government advocates believe that open access to public sector data can be a powerful force for public accountability, by making existing information easier to process, combine and analyse. OGD can then promote greater transparency, and allow a new level of public scrutiny that can increase public accountability.

This can raise the level of public trust and the perceived responsiveness of government actions. The Open Government Declaration "Open Government Partnership" (September 2011)[7] is considered to have established the use of new technologies – information and communication technologies (ICTs) in particular – to spur data sharing in the context of political accountability. This then blurs the distinction between the technology of open data and the politics of open government. However, it is important to underline that open government and open data can each exist without the other. A government can be open, in the sense of being transparent, even if it does not embrace new technology, and a government can provide open data and still remain deeply opaque and unaccountable (Robinson and Yu, 2012). Making public sector data available in machine readable format indeed has the potential to improve service delivery and citizens' quality of life, but it may have little impact on political accountability. Additional measures for enhancing government accountability and transparency, as well as democratic control, may be needed in addition to open access to public sector data.

Self-empowerment, participation and engagement of citizens

Another point often made by open government advocates is that opening government data enables individuals to make better decisions in their lives and increases participation in public affairs. Normally, e-participation is part of a government's broader digital government policy. It is the element aimed at harnessing IT use for openness, transparency and collaboration within the public sector, but also at enhancing citizens' engagement in public life, e.g. in lawmaking, policy making and service design and delivery. OGD initiatives, particularly as they are supported by Web 2.0[8] and social media applications, are creating architectures for participation that enable citizens to be not just passive consumers of public sector content and services, but also active contributors and designers in their own right. The expanding use of new technologies, combined with the rise of the OGD movement, is seen as a key enabler and driver of self-empowerment, higher e-participation, and the public engagement of citizens.

Legitimate stakeholders are for example invited more openly into a participative and empowering relationship with government in terms of:

- working arrangements of the public sector and public governance more widely
- planning and land use issues

- service design and delivery
- community building
- dispute and conflict resolution and broader public policy and decision making as part of the overall democratic process.

Open access to public sector data, but also exclusive access to citizens' own personal data (i.e. "smart disclosure"), empower citizens to make more informed decisions that can enhance the quality of their lives (Howard, 2012b).[9] For this to happen, governments need to enable users to have access to their own data and decide how to use it (e.g. the Blue Button Initiative in the United States to give veterans complete control of their personal health records held by the public sector; or the Green Button, also in the United States, which is a similar initiative around individuals' energy use data).

It is equally important to empower the public sector workforce. Opening up government data can enable civil servants, many of whom are front-line professionals, to participate in ensuring that government is open and participative, and to develop applications that better respond to users' needs. Many civil servants see the real-time performance and impact of public services and public policies on citizens. Empowering them could generate appropriate data and other inputs that could in turn improve the service experience if they were given the data, tools and incentives to do so – for example by being enabled to participate in a professional capacity in citizens' social networks, offering advice and knowledge.

Moreover, many civil servants see a blurring of their personal and professional lives in terms of the tools they use; both could improve through a two-way exchange of experience and skills. Sensible structures are needed to ensure that civil servants are empowered this way while maintaining impartiality and a position of trust, from the government itself as well as from citizens. This requires also that civil servants be equipped with the necessary skills, tools and mechanisms (Millard, 2012) and guidelines.

But for this to happen, strategies and programmes are needed to build the next generation of civil servants. New skills are required, not only for IT but also for data science; predictive analytics to identify patterns and create models; a better knowledge of how to use Web 2.0 technologies for social engagement and to negotiate with and connect to people; and a finer understanding of emerging problems and use of IT use to solve them (e.g. cybercrime investigation).

Fostering data-driven innovation in the private sector

Increasing efficiency and effectiveness in public services delivery

Granting the private sector better access to public sector data can increase efficiency, effectiveness and innovation in public service delivery. The strategy is to provide innovators from outside governments with the opportunity to develop modular services that are more agile and targeted to citizens' needs than those developed in-house by governments (see Box 10.3). Even though the release of data online can raise a number of substantive enquiries in terms of government activities, from a public service delivery perspective its reuse can also lead to a significant reduction in the questions routinely received by public authorities, thus decreasing workload and costs. Additionally, the remaining questions concerning service delivery per se would be easier for civil servants to answer, as it would be clearer where all the relevant information could be found.

Box 10.3. **Countries releasing PSI to the private sector**

The Cultural Heritage Agency of the Netherlands is actively releasing their data and collaborating with amateur historical societies and groups such as the Wikimedia Foundation in order to execute their own tasks more effectively. This can result in improvements in the quality of data and ultimately make government departments leaner, while encouraging external inputs and new sources of knowledge, possibly making the departments more innovative. In addition, one could argue that the co-development of knowledge in this example increases not just the quality but also the awareness of the Dutch public authority's work, thereby further increasing its value and relevance.

Similarly, in *France* the new version of the French national Open Data Portal (www.gouv.data.fr) enables non-institutional actors to upload data collected or produced by the government that can be mashed and linked with data uploaded by the public authorities. As highlighted in Chapter 4, this can lead to the development of innovative products such as apps, and to greater public-private collaboration in jointly identifying and developing solutions to problems. The government's credibility and accountability are ensured, as only the data provided by the French Government are released as certified open government data.

With the same aim, the government of *Canada* has committed to creating an open data institute (the Canadian Open Data Exchange, or CODX) as a national marketplace for those engaged in the commercialisation of open data, and will among other things allow the development of new tools and applications that access and manipulate public sector information; establish a framework for open data standards; and include the articulation of industry standards for presenting and providing access to open data for key sectors.

Enabling new goods and services in the private sector

As the importance of data in the development of new services, products and markets has increased dramatically (Koski, 2011), open access to public sector data can stimulate innovation in the course of that development. When public sector data are open, however, access to the data per se no longer provides a competitive advantage to firms with exclusive data access agreements. Competitive advantage has to come from offering innovative value-added services on top of data, and providing opportunities for business start-ups. The private sector (technology developers) is expected to be among the primary users of public sector data sets in pursuing their commercial exploitation. A profit incentive can help to drive innovation and experimentation; one would expect the best ideas to be emulated and improved upon, as no service provider has the monopoly on data.

There is in particular cross-country evidence that significant firm-level benefits are to be had from free or marginal cost pricing, with small and medium-sized enterprises (SMEs) benefiting most from less expensive data and the switch to marginal cost pricing (Koski, 2011). For example, analysis of 14 000 firms in architectural and engineering activities and related technical consultancy services in 15 countries in the 2000-07 period shows that in countries where public sector agencies provide fundamental geographical information for free or at maximum marginal cost, firms grew about 15% more per annum compared with countries where public sector geographic data have cost-recovery pricing. Positive growth comes one year after switching to marginal cost pricing, but growth is higher with a two-year time lag. Apart from SMEs (once again) benefiting most from cheaper geographical information, switching to marginal cost pricing of PSI substantially lowers SME barriers to enter new product and service markets.

Public sector geographic data also have the potential to enhance transportation and environmental performance. The value of improved time allocation can be estimated from data for Norway, where a minimum of two hours per citizen per year could be saved through better access to public information (Norway, 2013). A simple GDP-based pro-rata calculation for the OECD gives USD 6.4 billion in annual value of individual time saved if better access to public information saved only two hours' time per citizen per year. Furthermore, European Law requires environmental impact assessments and strategic environmental assessments. The European assessment market has been estimated to be valued at EUR 1 billion per year for national assessments (Craglia et al., 2010); improving accessibility of the information required could save up to EUR 200 million per year for these assessments. Including sub-national assessments values could be 10 times higher, i.e. a market value of EUR 10 billion, with potential savings from better information of EUR 2 billion across the EU27 countries. Further initiatives such as GovLab in the United States (see Box 10.4) are under way to study the potential of public sector data for businesses.

Box 10.4. **Open Data 500**

The Governance Lab at New York University (the GovLab) undertook a comprehensive study of US-based companies that use open government data to generate new business and develop new products and services. The objectives of the Open Data 500 are to:

- provide a basis for assessing the value of open government data
- encourage the development of new open data companies
- foster a dialogue between government and businesses on how government data can be made more useful.

Having launched the website OpenData500.com with in-depth information on 500 companies in early April 2014, GovLab is now focused on organising roundtables with the aim of spurring interaction between government agencies and their stakeholders to accelerate and improve the release and use of valuable open government data. The dialogue should help prioritise the release of open data sets for businesses and developing ongoing collaboration and feedback loops from data users to providers. Initial analysis of the data collected through the survey filled in by the 500 companies led to the identification of 13 main types of companies using OGD; the main types are data/technology, finance and investment, business and legal services, governance, health care, logistics and transportation, research and consulting, and energy.[10] Initiatives such as Open Data 500 are key to fostering the development of an ecosystem in which data providers improve their knowledge of data users' needs, which can help them make their open data programmes more effective.

Source: The GovLab, 2014.

Estimating the wider impact on the economy

The approximate size of the OECD market for PSI and the broader economic impacts of PSI are estimated in this section (see Vickery, 2011, 2012 for the approach and references). The results presented here are based on using aggregate studies available to estimate plausible values for the PSI market, potential gains from freeing up access, and wider economic impacts that could accrue from using PSI across the economy. Further estimates could be provided if relevant aggregate studies are available from (for example) Canada, Chile, Israel, Japan, Korea, Mexico, the United States, or key partner economies.

Market size and aggregate economic impacts of public sector data at the country level are available for Australian spatial data-related economic activities, with results generated from a general equilibrium model of the Australian economy (ACIL Tasman, 2008). In the Netherlands, similar estimates are available of the size of the geo-information sector (Castelein, Bregt and Pluijmers, 2010). Productivity-related impacts on the New Zealand economy from the use and reuse of spatial information have been estimated using a general equilibrium model. Benefits from removing barriers to use, improving infrastructure, and expanding training are also estimated (ACIL Tasman, 2009). For the United Kingdom, estimates of gains from opening up access to digital, non-personal, public sector information are also available (Pollock, 2010).

OECD values are derived by prorating available national data to give estimates for Total OECD using macro data from OECD (2014b) and available EUROSTAT data on the European Union economy.[11] The same method was applied using national and OECD data for: a) GDP shares, b) computer services spending, and c) ICT spending by government (WITSA, 2009) for each set of national data. The three sets of results for each set of national data were pooled and the mean calculated. In the case of estimates based on geospatial data, it is assumed that the geospatial market/impact is about one-half of the total PSI-related market/impact,[12] and that one-half of the PSI-related market/impact comes from government PSI. Both assumptions are conservative. Geospatial information may be considerably less than one-half of all PSI, and governments are the basic source of information for probably more than one-half of all PSI-like activities. Furthermore, estimated values within and across different sources were reasonably comparable, suggesting that the averages provide reasonable albeit low estimates of the economic features of PSI markets and the impacts of PSI use.

Averaging the OECD PSI market estimate derived from Netherlands data (USD 113 billion) with the estimate from Australian data (USD 82 billion) gives an estimated OECD PSI market of around USD 97 billion in 2008.[13] Various studies have reported PSI market growth rates in the range of 6-18% per year (Castelein, Bregt and Pluijmers, 2010; Coote and Smart, 2010; Fornefeld, 2011; MICUS, 2010, 2009). Taking 7% per year as a lower estimate, the OECD PSI market would have grown to around USD 111 billion by 2010 provided that it continued earlier growth and was not dramatically affected by the recession. This value is estimated in the same way as, and is comparable with, the estimated EU27 market of EUR 32 billion in 2010.

Averaging the OECD estimate derived from Australian data (USD 557.5 billion) with the estimate derived from New Zealand data (USD 461 billion) gives estimated OECD aggregate economic impacts of around USD 509 billion in 2008. There could be approximately USD 194 billion of additional gains if barriers were removed and the data infrastructure improved, as described in the New Zealand study. That is, if PSI were opened up, skills barriers removed and the infrastructure more effective, aggregate direct and indirect economic benefits for OECD economies could have been of the order of USD 700 billion (1.7% of GDP) in 2008, and more in 2010.

United Kingdom estimates were used to give an approximate value of annual gains from moving from an average cost/cost recovery pricing model to marginal cost pricing for digital public sector information (Pollock, 2010). Upper range values for the OECD are estimated to be USD 127.9 billion to USD 170.6 billion in 2009, or alternatively USD 45.5 billion to USD 56.9 billion for middle range estimates. These ranges assume that the structure of public sector information and related markets and pricing models across the OECD area are similar to those of the United Kingdom (average cost/cost recovery pricing in many cases). From

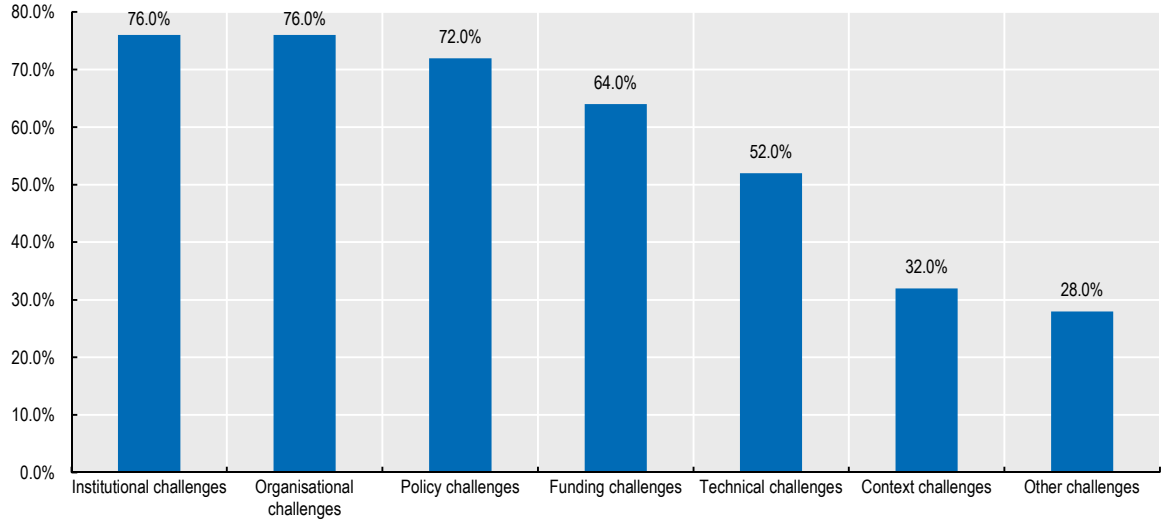
the upper range OECD welfare gains of USD 127.9 billion to USD 170.6 billion, a value of USD 145 billion is adopted in OECD, 2015.

10.2. Key challenges in implementing open data strategies

Initiatives to enhance open access to public sector data can be undermined by problems related to implementation, organisation, technical challenges and administrative delays, as well as those due to existing legal obstacles. If not properly tackled, implementation challenges might obstruct or restrict the capturing of benefits of national efforts aimed at spurring DDI based on public sector data. Technological, legal and financial restrictions, among others, may limit data access and reusability (e.g. making it difficult to fund data or find valuable ways to reuse data). Addressing various challenges related to technology, financing, organisation, culture, policy, and legal frameworks is essential to create an ecosystem, and build sustainable business models for PSI and OGD initiatives that can bear the desired fruit.

The most important challenges to furthering the development of OGD initiatives relate to policy, and funding challenges are most commonly cited as the second most important. The results shown in Figure 10.4 underline that the main obstacles for implementation of open data in governments are not technical but are linked to legal barriers or resistance within organisations.

Figure 10.4. **Open government data's main challenges as reported by countries**



Note: Other challenges include: cultural challenge both in government and in society about information and data management (Mexico), Multi-jurisdictional challenges, i.e. ensuring consistency in the environment within which OGD is being implemented (Canada), interoperability (Portugal), cultural change within the administration (Germany), lack of evidence of impact making selling the agenda to departments difficult (United Kingdom), demand-supply balance (Denmark).

Source: OECD survey on Open Government Data, version 1.0, 19 April 2013.

Policy challenges

Disclosure policies may limit data transparency and cause lack of clarity regarding who owns public sector data, and in so doing restrict the right of the public to use the data. In some cases, for example, public sector information or data are sold or come with restrictive copyright licences that prevent reuse. This may cause an unresolved conflict

between the right to access information as an inherent part of the right to freedom of expression, and limitations on reuse from copyrights or charges for commercial use.

The lack of procedures and standards on how to deal with open data in governments (e.g. lack of tools available to make data open, lack of validation structures and guidelines, lack of guidelines on data collection) can compromise the quality of the data and eventually the output of OGD initiatives. The adoption of an overall national strategy for PSI or OGD can help overcome many of the issues highlighted above. Not only can such a strategy clarify matters pertaining to licensing and standards, but it can also define a national approach and targeted goals that can help guide and structure actions and initiatives at all levels of government.

Technological challenges

Public sector data often are not harmonised given that individual units collect and/or produces their own set of data using different metadata, formats and standards. This can make it difficult from the user perspective to know which piece of data is valid or should be trusted. Critical to access is to know the source of what one is searching for, and in many instances where to start searching is a challenge. Accessibility can also be limited if data cannot be reused, and data transparency may be hindered if data are not simple to access or reuse due to their format. Additional technology-related shortcomings include the need to: i) improve information technology infrastructure, ii) enhance privacy and information security, and iii) integrate open data tools and applications.

A second layer of technical challenges can emerge when the federal government seeks to impose co-ordination or consistency across the broad range of rulemaking processes, data and portals enabling access to public sector data. Even though the establishment of a single OGD portal should not be the goal – and is far from being the best advisable solution for implementing OGD – a single point of access to government open data can certainly ensure integration of shared data input from various sectors of government, and can greatly enhance accessibility (see Box 10.5). Therefore, a lot of emphasis is often placed on the establishment of a single portal. Most OECD countries have indeed developed an online OGD centralised platform with the idea of increasing citizens' and private actors' access to a growing variety of government information made available as open data. However, to meet government-wide needs in terms of data management, any decision to create a single portal should be developed through a collaborative approach, to create ownership and secure sustainability. The trade-off between standardisation and experimentation, and concerns about incomplete or inaccurate data in centralised government repositories, are difficult problems that most OECD member countries are currently dealing with.

Box 10.5. **The case of Regulations.gov**

Regulations.gov is a government-wide docket publishing system created in the United States in response to the E-government Act of 2002, and launched in 2003. It is used today by most US departments and agencies (Regulations.gov, n.d.). The policy of the Office of Management and Budget (OMB) not only requires its use but also precludes the agencies from using "ancillary and duplicative" docketing and rulemaking systems of their own design (OMB, 2004). This exclusivity rule, combined with the difficult interagency politics involved in honing system features, is considered by many to have led to a bare-bones approach that leaves out the agency-tailored functionality found in many of the systems it replaced.

Concerns about cost sharing have also led the system to omit even features whose usefulness and desirability are a matter of broad consensus (Farina et al., 2008). Regulations.gov was launched with a limited search engine and no browsing capability, so that only those who already knew the terms used to categorise rulemaking documents were able to use it effectively. Five years later, a relaunched version of the site offered up its limited inventory of computer-readable data directly to the public (in this case, using a single rich site summary (RSS) feed, which allowed any interested person or group to create an alternative, enhanced version of the website. This has permitted the creation of OpenRegulations.org, which competes with Regulations.gov by offering "paired [sic] down, simple-to-navigate listings of new agency dockets" and a more sensible set of RSS feeds, one for each individual agency.

A recent OECD (2015) survey of government strategies to enhance the reuse of PSI highlights that all countries are aiming to achieve machine-readability and interoperability among data sets. They also hope to switch to or encourage the use of open standards (see Table 10.1). However, the reality is at some distance from achievement of these aims, and varies considerably across countries and features. An Australian survey on PSI management across 191 government agencies showed that 38% of them reporting that all or most of their PSI is in open and standards-based formats, and 58% reported routinely applying metadata to information published online (OAIC, 2013). In addition, at the end of 2011-12, 90% of the Australian National Library's collection was catalogued and searchable online (survey reply, Australia).

While new material is often provided in machine readable formats, older material generally is not. The response of the United Kingdom, for example, pointed out that a great deal of previously saved information is locked in PDFs or other unprocessable formats, and not in linked data formats. Similarly, not all PSI material on central government portals is available in open standards. This is the situation in most countries but not all, due to the evolution of such standards over time and their relatively recent widespread diffusion and use. Metadata are also less widely associated with data sets than might be hoped.[14]

Table 10.1. **Machine-readability, open formats and interoperability**

Country	Machine-readable	Open source / standards used	Metadata available
Australia	Data searchable	Where possible	Available
Belgium	Minority of data	Minority of data	
Canada	Large proportion of data		Common profile
Chile	Yes in principle	Work in progress	No. Technical guide being developed
Czech Republic	Data provided in formats of creation	Unrestricted use	
Denmark	Variable, depends on subject area	Variable, depends on subject area	Variable, depends on subject area
Estonia	Varies greatly. Information Society Strategy 2020 concentrates on making public data available in better machine readable formats. Green paper on machine-processable formats planned for 2014	Use of open formats is moderate or poor	Availability of metadata is moderate or poor
Finland	No reliable information	Planned	Planned, international standards
Hungary	Preferred for PSI. Not a requirement for freedom of information		Metadata database available for centralised public data portal
Japan	Planned. Significant amount machine readable for statistics	Significant amount of open format data for statistics	Provided via registration on data catalogue site
Korea	Significant portion of open data are machine readable and released, in principle, in machine readable format		Metadata are available for data registered at data.go.kr and further metadata will be available systematically
Mexico	Working on it via the Federal Public Administration's Interoperability and Open Data Scheme	Working on it	Available for an increasing set of statistical databases
Portugal		All, on national data portal	Most. Working towards mandatory availability
Slovak Republic	Standardised, but wide variety		
Slovenia	No express provision	Actively promoted	
Spain	Important part		Minimal already; Standardisation planned
Sweden	No general information		
Switzerland	Planned. International compatibility	Planned. International compatibility	Planned. International compatibility
United Kingdom	Recent data are machine readable		
European Commission	Source data yes	Not always	Catalogue metadata available

Source: OECD (2015) review of the OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information.

Economic and financial challenges

Economic and financial challenges are hindering fast-paced development of PSI and OGD initiatives in several OECD countries. In particular, many governments wish to recover costs, partly for budgetary reasons and partly on the grounds that those who benefit should pay. However, the calculation of the overall benefits can be problematic. Moreover, as Stiglitz *et al.* (2000) have argued, if government provision of a data-related service is a valid role, generating revenue from that service is not. There are further key aspects that need to be taken into consideration; these are highlighted in the paragraphs below.

Revisiting financing and costing models

The common assumption that making data available as open data is just a product of what happens already inside the public sector and therefore does require new investments is not entirely correct. Open does not imply free of costs, as there are some potential costs that need to be considered when making data openly available. There is, for instance,

substantial commitment and investment on the part of agencies as they need to acquire new skills, train employees, purchase technologies, and upgrade network infrastructure. There are indeed human-resource costs associated with ensuring timely updating of data as well as with organising and preparing information to be put on line – particularly if the decision is taken to develop a special portal that may require an IT and design team. The additional costs for timely data publication, or coherent production of high-quality data, are normally held by each agency.

In addition, converting large volumes of data into reusable formats can have significant cost implications, particularly if there is a high level of proprietary software use. Initiatives such as converting government data to semantic web and linked data formats, as well as enabling partial access to large volumes of data (through e.g. anonymisation[15]), can be time-consuming and therefore costly. Because of these additional costs there is some reticence on the part of government bodies, which can result in refusal of even partial access to a requested database, even if privacy concerns are eased. However, to comply with the right of access to information, public bodies often have no option but to take the time to remove the sensitive data and then grant access. This has a cost that needs to be factored into the overall cost and benefit analysis of enhanced access to public sector data.

Countries such as the Netherlands and Denmark are looking into developing a business case with funding and alternative financial models to augment investments on open data and ensure the buy-in of public sector agencies. This approach emerges also from the need to compensate the loss of revenues claimed by many agencies as consequence of the abolition of fees within the new open data regime. Especially in times of austerity, governments are concerned by the cost of opening up public sector data; worries are worsened by the fact that such costs – such as in the case of data production – have not been sufficiently appraised so far.

As highlighted in OECD (2015), sales of PSI (including public sector data) generate very little direct revenue for most governments compared to their costs (around 1% of expenditures to make the data available). The notable exceptions are found in the Netherlands and the United Kingdom, and even in these countries sales are a maximum of around one-fifth of expenditure incurred by the agencies generating the information or data. In contrast, the benefits for society – including for the private sector – can be significant (as highlighted above), and can lead to additional tax revenues from downstream private sector activities.

In terms of the balance between revenues forgone and benefits from free access, a Danish study, for example, explored the impacts of making address data free (Danish Enterprise and Construction Authority, 2010). Official address data have been free of charge since 2002. The study showed that direct financial benefits for society in the period 2005-09 were around EUR 62 million (USD 83 million), while total costs were around EUR 2 million (USD 2.7 million). In 2010, estimated social benefits were around EUR 14 million (USD 18.8 million), with costs around EUR 0.2 million (USD 0.27 million), with 30% of the benefits in the public sector and 70% in the private sector. The study only included the direct financial benefits for the 1 200+ parties receiving address data from a public data server -distributor; not included were additional economic benefits in later parts of the distribution chain, for example in GPS systems. Further benefits could be expected if the availability of official addresses is extended to business registration addresses and utilities.

In Finland the Ministry of Finance reviewed the 2009-10 income of key governmental agencies from information disclosures/sales (survey reply, Finland). Income was estimated at around EUR 30 million (USD 40 million) per year from the private sector. As Finland progressively shifts to an open data strategy, adjustment for this income is being reviewed on a case-by-case basis. In Switzerland, many federal offices provide their data for free; nevertheless, federal revenue from that provision was CHF 41 million (USD 44.6 million) in 2012 (Federal Department of the Interior, 2013). The Swiss study produced estimates for the federal administration of the overall balance of free data between revenue foregone, new tax revenue, efficiency gains and switching costs. Annual net direct benefits were estimated in the range CHF 2.9 million to CHF 20.3 million (USD 3.2 million to USD 22.1 million) over three years. It was concluded that Switzerland would benefit from introduction of open government data (open PSI). The Swiss federal administration would obtain clear efficiency gains, provided the issue of compensation for federal offices can be settled.

In addition, a recent OECD (2015) survey of PSI strategies suggests that countries have not had particular difficulties in funding the switch to free and open data and information, and that this has not been the major barrier that was foreseen in the past (see Table 10.2). Half of the respondents (12 of 20 countries plus the European Commission, including countries reporting both) did not have special funding or budgets for the switch to open and free PSI strategies. The sources of finance were largely internal, or derived from reallocation of existing funds. The United Kingdom did not foresee significant increases in spending, and the European Commission foresaw lower administrative expenditures from switching to open strategies. For those countries where special funding was envisaged, it came from either within the budget process (Chile, Denmark, Estonia, Finland, Japan, Korea, Mexico) or from broader funding packages for modernisation or open government (Portugal, Slovenia).

Table 10.2. **Budgeting for the costs of opening up public sector information**

Country	Special funding	Sources of funds	Issues, other
Australia	No	Included in existing budgets (but central funding for central government portal and support to whole of government)	Agencies responsible for own licensing practices
Belgium	No		Study under way on budget models
Canada	No	Included in existing budgets (but central funding for central government portal)[1]	
Chile	Yes	Budget includes transparency funding	
Czech Republic	No	In overall budget	
Denmark	No/yes	Good Basic Data for Everyone resources provided at central, regional, local levels	
Estonia	No/yes	Resources inside normal general budgets. Ministry of Economic Affairs and Communications has additional central funding to accelerate open data projects for other ministries, agencies and local governments	
Finland	Yes	Decisions part of budget process (plus funds for national open data programme)[1]	Stepwise introduction of opening data
Hungary	No	No specific budget funds	
Japan	Yes	Budget funds allocated 2013 fiscal year, adjusted for 2014	
Korea	Yes	The Ministry of Security and Public Administration allocates budget for pan-government efforts and each ministry/agency allocates relevant budget	
Mexico	Yes	Budget funds allocated to the Federal Institute for Access to Information and Data Protection	Over half of Institute funds promote information access
Norway	Yes	Central government for central open data activities[1]	
Portugal	Yes	Part of Global Strategic Plan for Rationalisation of ICT Costs in Public Administration (PGETIC)	Funded within overall PGETIC envelope
Slovak Republic	No	No extra funds provided	
Slovenia	Yes	Part of Open Government Strategy. Special funds planned for opening PSI	
Spain	No	Internally financed	Small budget to facilitate opening
Sweden	No		
Switzerland	Yes	In planning stage	Revenue loss compensated
United Kingdom	No	Significant increases in spending on national data strategy not foreseen nor additional administrative complexity (but financing e.g. the Open Data Institute and aiding departments release their data)[1]	Aim to broaden objectives and sharpen planning and controls
European Commission	No	Included in budget	Free reuse policy lowers administrative expenditures

1. Information from the OECD survey on Open Government Data.

Source: OECD (2015) review of the OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information.

Nevertheless, several respondents pointed out that in times of budget pressures and cuts in government expenditures, it is important to articulate clearly the advantages of opening up public data for wider use and, where necessary, to compensate the providers of public sector data for any initial extra funding necessary to open up and digitise the data (see the following section). The 2013 OECD Open Government Data survey reports that no government has adopted a methodology to measure returns on investment in OGD, and that there are relatively few and only scattered attempts to track economic or social gains from the reuse of OGD. Nine out of twenty-five countries reported that they are working in this area, mainly in terms of developing and collecting case studies. And even fewer countries have information, for example, on government income or the value of extra tax revenue from new business associated with the commercial exploitation of

public sector data. This highlights the need to establish a clear measure of the potential costs and benefits of opening up public sector data. Doing so could help governments build a better business case for open access to public sector data. A clear business case could in turn help secure the needed political support and facilitate implementation in pragmatic and affordable ways so as to avoid unnecessary burdens and loss of revenues.

Establishing the right ecosystem

It has been suggested that one solution, for the longer term, is to design databases with the right of public access in mind – which appears to be increasingly easy, at least from a technical perspective. It is possible, for example, to build a database that performs one-way encryption. This permits e-mail addresses to be included in a database, but in another table that is linked via a hash value so that when the data are shared, the e-mail addresses can be separated. Similarly, there are many solutions to releasing information that come at a very low cost, and it would be advisable to see these as part of the day-to-day activity of public bodies, such as posting full data sets in open source formats on government websites, properly tagged with metadata so that the information can be found but with no other special formatting or presentation.

The "business model" for OGD also needs to take into account where potential benefits may accrue, and how to align funding and incentives. When government provides reusable data, the practical costs of reuse, adaptation, and innovation by third parties are significantly reduced. It is reasonable to expect that the low costs of entry will lead to a flourishing of third party sites extending and enhancing government data in a range of areas – rulemaking, procurement and registered intellectual property, for example. This approach could be adopted by those governments that decide to shift their online focus from developing finished websites based on public sector data to the infrastructure that allows new sites to be created. If the creation of infrastructure causes better third party alternatives to emerge, then the government entity can cut costs by limiting its own. This reinforces arguments in favour of better appraising the costs and benefits of OGD and PSI, as well as a clear strategy that provides incentives to public officials to invest in related activities. Such an approach would more clearly frame co-ordinated and efficient decisions on government IT and information architecture, and could secure alignment with wider government IT procurement strategy.

If on the other hand third party alternatives to the government site do not satisfactorily emerge, then the public site can be maintained. The overall picture is that government IT costs will decline in those areas where private actors have the greatest interest in helping to leverage the underlying data, while government IT costs will increase in those areas where, for whatever reason, there is no private actor willing to step forward and create a compelling web service based on the data. Governments are keen to collect evidence on recent initiatives showing that putting raw data on line demonstrates that it can be considerably cheaper than presenting the data to the user via a custom web interface.

Organisational challenges

Implementation of OGD requires also dealing with a number of organisational challenges, described in the paragraphs below.

Ensuring accountability, quality of data and responsibility in a context of collaboration

Given the complexity and crosscutting nature of public sector data, governments need to establish the appropriate institutional structures. Tasking a government body – often the centre of government (e.g. the prime minister's office) – with championing, co-ordinating and providing support for and leadership of open government data initiatives and programmes has been seen as a way to bring the various stakeholders on board. Having a ministry (dedicated body) in charge of soliciting from governments the various data sets that will then be made public has been additionally considered as a way to ensure timely and full compliance with the national strategic directions. This dedicated body can sustain collective work to strengthen data integration across different parts of the public sector, help build better capacities across governments to deal with emerging concerns (e.g. privacy/transparency), and ensure that those making decisions about the release of data do so in a rigorous and consistent fashion.

Empowering independent oversight bodies to demand and to publish information on budgets, procurement and expenditures is considered crucial for ensuring data transparency. Several countries, e.g. the United Kingdom, are considering establishing independent ethics and governance groups to oversee policies and procedures for improving the use of administrative data. In addition, some countries have assigned the role of open data "evangelists" to a person responsible for promoting open data across the public sector (see Box 10.6).

Box 10.6. **The Evangelist for US Data.gov**

On 10 August 2013 a position was posted for an "Evangelist for Data.gov Open Government". The job description indicated that the candidate for the role was required to show four very different capabilities: 1) extensive outreach and communications skill and experience; 2) extensive experience in designing and implementing open government systems; 3) a proved research record for identifying and developing new technologies; and 4) managing a complex data and information environment that encompasses data ranging from public to classified. The job description also indicated that the Evangelist would have to work extensively with multiple parts of the government, thus underlining the importance of understanding the multitude of policy issues inherent in the release of information key to Data.gov. Hence, the role required knowledge of, and access to, an extensive network of people, organisations and experience, given the many linked areas of public outreach and engagement. The role was established also to spur knowledge dissemination and "evangelisation" in relation to the development and use of Data.gov, with the goal of gaining the greater involvement of agencies and other stakeholders such as the open government community and the mash-up programmer communities. The announcement clarified that the Evangelist was expected to create excitement and drive around the programme to facilitate practical field application of leading-edge technology issues with important stakeholders.

Source: Federal Business Opportunities, www.fbo.gov, accessed 15 May 2015.

Ensuring sustainable change through the data ecosystem

Creation of the right ecosystem is essential, not only to reap the economic benefits but also to generate the value of OGD initiatives, in social and political terms. As indicated earlier in this chapter, data used by third parties, as well as the use of the apps developed based on them, are absolutely necessary to make public sector data sustainable for value

creation. "Asthmapolis" in the United States is an excellent example of an app developed thanks to the ecosystem around public sector data, and which has brought social value and improved quality of life to a vulnerable segment of the population: people with asthma. Public data and data provided by people affected by the disease have been merged in the app to enable identification of highly dangerous spots for asthmatic people in the United States. Hospitals have recorded a decrease of 25% in incidents since the app was created.

However, establishing the right ecosystem around public sector data is no simple task. It requires the involvement of all actors and provision of the right business case to spur usage. It also entails identification of the various categories of actors; adoption of policies built around issues that are universal; nurturing of a culture of public sector interaction with the actors; and reaching out to some that might normally be less actively involved in public affairs (e.g. civil society organisations operating in geographically remote areas, and as such more aware of data that might be needed to develop target services that would better serve the local community). At least three roles of actors identified in Chapter 2 can be highlighted here as highly relevant for the use of public sector data:

1. *Data (service) providers* (i.e. in the public sector, academia, media and private sector).
2. *Data-driven entrepreneurs* (i.e. media, developers, civil society) – which essentially provide products to make sense of and create value out of public sector data. Media, for instance, can tell interesting stories based on such data; developers can produce web services and apps; civil society organisations can spot the relevance of certain data for specific segments of the population (e.g. charities in remote areas), play a critical role in building capacities at the community level, and create a culture that appreciates the relevance of public sector data.
3. *Users/citizens* – Communities need to use public sector data and engage to get the most out of OGD initiatives. Libraries also play a key role in relation to data mining and as facilitator of accessibility to data, particularly in countries' remote areas, and thus enhance the cost-effectiveness of access.

Interaction among all actors is essential (see Chapter 2). Knowing and understanding each category is important, as it helps grasp what value can be created for the community and how this can be achieved. The key questions are, for instance: who are the main members of the user community? Who leads interaction with them, what are the expected outcomes of this interaction, and how can these be measured?

Good examples of strongly collaborative ecosystems exist at local government level (see Chapter 9). The City of San Francisco, for instance, is characterised by a culture based on a strong sense of community, with a relatively large number of citizens and ICT activists forming a dynamic ecosystem supporting a strong bottom-up innovative context. San Francisco can also count on the open-minded and collaborative attitude of the city authorities as a real driver for OGD. And in that, San Francisco is not unique. It indeed presents many elements that typify several OECD medium-sized cities, as well as large municipalities. A way to replicate the positive experience might be to adopt a strategy that leverages these conditions where they exist, or fosters their development where they are lacking. Establishing collaboration frameworks may also help to ensure the involvement of different actors (e.g. SMEs that may be important incubators for innovation but that are as yet little aware of opportunities generated by OGD.

Engaging with the wider community in a two-way conversation to build capacities and find agile solutions

Churning out data is not sufficient to create value. Robust engagement models also need to be in place, to allow two-way dialogue to take place between the public sector and the users of public sector data (i.e. individual developers, SMEs, citizens, civil society organisations, academics and private companies). It is key for governments to (e.g.) focus on users' needs and for users to (e.g.) provide feedback on the data sets they would like to see released as a priority. Capturing feedback may result in value creation, as doing so enables new features, new lines of business, new markets, new competencies, new services and new tools. Similarly, users can spot anomalies and mistakes in government data and thus contribute to improving public service delivery and policy making. Developers at the cutting edge of technology can be kept up to date on new data sets being released, and governments can find help in doing things differently and in more agile ways.

The government of the United Kingdom is for instance working on a Government Developer Engagement Strategy, setting out principles for how individual government departments are expected to engage with the development community. Several governments' initiatives launched competitions with the intention to encourage reuse of public sector data (e.g. the Apps for Democracy, run for a 30-day period by the government of the United States – which apparently led to an estimated 4 000% return on investment – or the similar Finnish Apps4Finland. The Norwegian initiative Nettskap 2.0 resulted in the development of 135 apps). Other initiatives have fostered close collaboration between individual civil servants/public sector bodies and civil society. As an example, in the Netherlands the online network "civil servants 2.0" (Ambetnaar 2.0) was developed together with initiatives sustaining a community-based and collaborative approach, such as the running the data catalogue overheidsfeeds.nl or the event BarCamp on Open Government).

Alongside mobile technologies, social media can also play an important role in inspiring or enabling many OGD uses. This underlines the relevance of informing communities of practices to sustain OGD initiatives, and involving them to help create a network of actors. Social media channels can also help capture users' feedback and create a need for use, i.e. get the data to where people really need them. However, engaging users requires adequate sills and resources.

In order to ensure the views of open data users are captured, the United Kingdom has established a group in its Cabinet Office that comprises 14 officially selected volunteers from civil society and the private sector, who advise the government on the data it should release.

Revisiting internal processes to support data release workflows

Actual implementation of open government data portals requires adequate workflows for data gathering, integration, validation, release approval granting and reuse promotion. In some instances the process of online data release is supported by an organisational culture already oriented towards data sharing and reuse, which facilitates process re-engineering. In other cases the internal culture of the relevant public sector institution is not immediately conducive to data sharing, which requires additional efforts. All departments and ministries must commit to these efforts for the success of open data in the public sector (Box 10.7); for some, that may require a significant cultural challenge.

Box 10.7. **The UK open data white paper: Unleashing the potential**

In June 2012, the UK Cabinet published its open data white paper, which set out how the government intends to put data and transparency at the heart of public services. The document underlines the intention of the central government to facilitate access to public data; make it easier for data publishers to release data in standardised, open formats; and engrain a "presumption to publish" unless there are specific reasons not to do so (for instance relating to privacy or national security). These objectives are integral to the full commitment to make open data an effective engine of economic growth, social well-being, political accountability and public service improvement in the United Kingdom. In order to frame a feasible public sector implementation plan for open government data, the paper highlights that following two years of the centre of government leading the initiative, government departments are expected to take a greater role in driving efforts forward. Therefore, alongside the white paper, each government department published their first open data strategy. Each strategy contains a department's commitments for proactively publishing data over the next two years, which will complement their existing statutory publication schemes. These strategies represent an important step forward in the way the country is making data readily and systematically accessible; they are a core requirement of each department's activity.

Source: UK Cabinet Office, www.cabinetoffice.gov.uk.

Cultural challenges

Legislation, IT platforms and applications need to be matched by a culture within the public service that supports a presumption to publish, release and share data. The sections below underline some of the key cultural challenges that many governments around the world are still dealing with, within the public sector and in society at large.

Increasing public interest and engagement

Raising capacity relevant to OGD and awareness of civil servants, citizens, civil society organisations and the private sector with regard to their rights is important for society as a whole to fully capture the benefits of public sector data. Government departments, in partnership with civil society groups, can for instance create awareness of legislation and policies that empower citizens to access information, such as the Access to Information or Freedom of Information Acts. Additionally, undertaking research to establish users' information needs and barriers to information use and reuse, or seeking public-private partnerships to encourage data use to foster innovation, can lead to ventures for the worthwhile reuse and redistribution of and universal participation in OGD, such as application development and provision of e-government services.

Recognising the value of crowdsourcing

Of critical importance for governments is to recognise the value of crowdsourcing to find the "talent" outside the public sector that can use data, create value from it and exploit it (see Chapter 3 on the potential of crowdsourcing data analytic capacities). This is not necessarily easy, as successful crowdsourcing also depends on a sufficient scale and representativeness of participation to get valuable results. A critical new resource to fuel such changes is public sector data made available in machine readable data sets that can also be searched, manipulated and interlinked using freely available tools. To date there is still only a limited number of governments that have embarked down this path to

any real degree and even fewer local and regional governments, where the benefits are likely to be greatest. The United States federal government as well as some cities in that country, and the United Kingdom, Australia and France as well as a handful of other governments have been leading the way in this respect.[16] Companies and SMEs, including start-ups in some countries, are exploiting such data to expand business and create jobs, while a few governments are using the data to encourage innovation camps, "hackathon" events, and other competitions to create new services and insights for policy making.

Providing incentives and building new capabilities for a cultural shift in the public sector

Missing participatory and collaborative elements, incomplete data and the lack of raw data represent much more than technological challenges. Solving these matters requires a fundamental cultural change in the approach of public authorities: from disclosure to proactive and smart disclosure, and from provision of information to provision of data that abandons the notion of interpretational sovereignty. The belief that making data public dis-empowers public officials – or makes them more vulnerable, since they risk unveiling faults – can at times create an environment among civil servants, or even policy makers, that does not fully support implementation of OGD initiatives. In some public sectors, these initiatives are actually producing a negative behavioural impact on civil servants, who show unusual resistance to collecting data. Governments are for instance increasingly considering developing training or awareness-raising programmes to help change the attitudes of public officers with regard to making data available to the public and improving its sharing with peers. Many governments are realising that cultural and administrative barriers to data sharing can best be addressed through engaging with, and crowdsourcing the experiences of, civil servants working with data, both on the front line and in central governments.

Additionally, governments must have the capability to analyse, interpret and consume the outputs of data and analytics work intelligently. This includes the capacity to debate the meaning of data and find ways to use it in democratic debate, as well as the ability to support more targeted policy making and improved service delivery. This is only partly about cutting-edge IT and data science skills; it is also about ensuring that public sector managers and policy makers are confident in combining data with sound judgement, and are aware of the need to encourage the pursuit of the OGD agenda, possessed of strong ethics and integrity.

Furthermore, even though having a firm idea of what data are available is an essential step for any government's OGD strategy, most governments currently do not have a comprehensive overview of the data in their possession. The government of the United Kingdom, for instance, organised information engineering programmes that forced more than 100 000 authorities to re-engineer their records; these are considered to have been essential for the success of open data initiatives. However, the cultural context matters, and the forceful approach that may have worked to make OGD initiatives successful in one country may not have the same rate of success in another. Governments can, for example, increase attention and foster valuable reuse of data by identifying specific economic and societal problems that they wish to see solved, or by providing incentives to reusers.

Finally, special efforts have been made by many governments while developing data portals to encourage the use of linked data. Skills in and experience working with linked

data may as yet be limited, but advocates of linked data approaches believe they could revolutionise how data are accessed and utilised (Davis, 2010). There is however much that governments need to do to reach that point. Data.gov.uk appears to be one of the few OGD initiatives where links among the different data sets have been created (Kalampokis, Tambouris and Tarabanis, 2012). The following figure is an example of how the linking can be depicted. More specifically, data from the Department for Education describing schools are linked to data from the Office for National Statistics. The "joint point" of these data sets is the LocalLearning Skills Council (LLSC) that is responsible for the specific school.

Ensuring the support of all stakeholders

Initiating dialogue among various stakeholders about the importance of sharing information and its benefits with the public can help secure their participation and ensuring their support. Current and potential reuse initiatives by the private sector, civil society organisations and individuals can be publicised to increase awareness of the benefits of opening up data.

Legal challenges

The legal landscape surrounding data sharing and opening is complex.[17] Having a consistent legal framework in place is critical to facilitating PSI accessibility and reuse; to improve secure data sharing between public authorities and with the wider community to improve insights, results and impacts; and to inform better policy making. Fragmented and diverse legislation concerning privacy, the reuse of data and (sometimes) related fees (e.g. in Sweden and Germany[18]) can create confusion for end users. PSI and FOI (freedom of information) legislation, as well as clear licensing guidelines, are a cornerstone of open access to public sector data. Guidelines and handbooks are among the useful measures a government may choose to adopt in order to facilitate and co-ordinate the work of agencies in their transition towards open provision. These guidelines cover technical and legal issues, economics and communication strategies. Several countries are already working on the development of such guidelines (e.g. Norway) or have recently published them (e.g. Spain,[19] France[20] and Denmark[21]).

Several member countries have adopted legal and regulatory frameworks to ensure adequate support for open data (see Table 10.3). Some countries are reviewing existing frameworks, or developing new ones. Mexico, for example, is reviewing its Access to Information Law, while Spain recently adopted the Law on Transparency (Law No. 274/2013 of 26 November) (see OECD, 2014c). The law has a triple purpose: to increase and strengthen government transparency; to recognise and guarantee citizens' right of access to information; and to establish good governance obligations to be met by public officials as well as the legal consequences of non-compliance. The law does not fill an absolute vacuum but delves into what has been achieved so far, correcting deficiencies and creating a legal framework to grant certainty to citizens' rights. The law for instance establishes a number of obligations to sustain proactive dissemination of certain information without waiting for specific requests from citizens. This applies to institutional, organisational, planning, legal, economic, budgetary and statistical data. Additionally, the law broadly establishes the *right* of access to public information, which may be exercised without having to justify the request (OECD, 2015).

Table 10.3. **Public sector information licensing practices**

Country	Licence used on central portal	National model licence
Australia	Free of charge under CC Attribution Licence (CC BY). Other licences may be used	CC BY defined as the default model
Belgium	Developing new licensing models including one restriction-free model	Standard federal level licence since 2007
Canada	New Open Government Licence. Similar to CC BY	Yes
Chile	CC 3.0; GNU General public licence (GPL) for software; and Open Database Licence (ODbL)	
Czech Republic	Generally non-exclusive; exclusive only if indispensable and in public interest	
Denmark	Recommended national licence, similar to CC BY	Yes
Estonia	No exclusive licences. Most PSIs free of charge with no specific conditions for use or reuse. Specific non-discriminatory licence conditions in some areas	
Finland	Under development. CC 4.0 and CC0 based (CC0 has no rights reserved)	Planned
Hungary	PSI agreement required for reuse	
Japan	CC licence for trial version of national data catalogue site. Licence for full-scale site to be determined	
Korea	No national licence policy, but at data.go.kr, conditions for use are stated for specific data	
Mexico	No information available	
Norway	Open licences where attribution permitted	Norwegian Licence for Open Government Data is a standard optional licence
Portugal	Non-exclusive licences. Central portal CC "BY" 3.0	
Slovak Republic	No general policy. Open government portal ODbL 1.0	
Slovenia	CC encouraged	Guidelines available end-2013
Spain		National model licence
Sweden	Licences relatively rare	No
Switzerland	Unified solution not yet available	
United Kingdom	All public data to be released under same open licence	Developing "New Open Licence"
European Commission	Reuse provided source acknowledged. Disclaimer rather than formal licence	

Source: OECD (2015) review of the OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information.

Additionally, a number of countries have focused on changing the legal context to enhance the impact of open data on good governance value, i.e. transparency and accountability. In May 2014 the United States adopted an innovative open data-related norm, the US Digital Accountability and Transparency Act (better known as the Data Act), which is supposed to bring a greater level of transparency and accountability to federal spending information by ensuring that agencies use a common set of data standards and by putting accurate, timely information on line for public consumption (see Box 10.8).

Box 10.8. **The US Digital Accountability and Transparency Act ("Data Act" 2014)**

In May 2014 the United States Government signed the Digital Accountability and Transparency Act (the Data Act of 2014). The result of adopting what is referred to as the "open data law" is a new policy aiming to standardise spending data and publish it on a single website within three years. The act also demands that the website be updated with new information at least every quarter, if not every month. The new law requires federal agencies to account for every dollar they spend on a single website, in an easy-to-read format, and aims to help people identify duplication, waste and fraud. It will take several years before all of its components go into effect, but the result should be federal agencies using a standardised reporting method to disclose their expenditures in even greater detail than previously. This is expected to enhance the transparency of federal spending, which the US Government regards as a means to achieve greater accountability to the taxpayers. The information was not necessarily hidden from the public before, but it may have required working with each individual agency to find and decipher it.

Under the Data Act, all spending information will appear on USASpending.gov, and visitors will be able to search through and download it. Supporters hope that the site will facilitate better oversight of government spending, and identify waste that can be eliminated. The bill also gives the government the option to create a centre dedicated to analysing the newly sorted data, so as to provide even more effective oversight on spending and further improve spending efficiency and transparency.

While the bill does not dictate the exact standard, it does require that the government chose something that is already widely accepted and not dependent on a single platform. The standard must also be able to be continually updated — a requirement that may help prevent the government from falling into a situation where agencies are all reporting information separately and in their own manner.

Because of the differences in national legal contexts and the difficulties in tracking actual implementation, legal developments are almost impossible to compare. Even though it is difficult to say how they compare to the US Data Act, there are a number of innovations in the legal framework of different countries supporting financial openness that are worthwhile mentioning, even though they are not necessarily enshrined in one single law. Brazil is a long-time pioneer in the field. As a result of passing the Law of Fiscal Responsibility, federal government agencies have since 2004 been required to publish all of their financial data on line in machine readable formats and on a daily basis through the country's Transparency Portal. The website contains vast amount of detailed and up-to-date information on government revenues and expenditures, procurement processes, and federal transfers to municipalities, states and individuals. Budget lines have both the official and popular names of the initiatives, and as a result the website is widely used by the media, government officials and citizens. Reports using data from the website led into investigations on the alleged misuse of public funds, and ultimately to the resignation of a minister. Civil society also used information to reveal how taxpayers' money is spent in Brazil.

In the case of the United Kingdom, instead of passing a single law the British financial transparency regime is a mixture of codes of practices, policies, amendments to the FOI law and governmental experimentation. Also not a result of one single law, South Korea's Digital Budget and Accounting System (dBrain) is seen as another innovative approach in the area of financial openness. The portal contains real-time information on budget formulation and execution, data on procurement processes, and a participatory budgeting feature where the central government, local governments, public

institutions and the public jointly decide on the allocation of resources. As an extra feature, citizens can also report alleged misappropriation of government funds, and may even be awarded up to USD 30 000 if allegations are proved. Finally, a recent law passed in Italy states that the information contained in SIOPE[22] will soon be accessible to the public in open data formats.

Notwithstanding these important developments, below are some important aspects and issues concerning the legal debate around open data in governments that remain unresolved:

- *The scope of right of access to information* – In principle, the right of access to information applies to all information held by public bodies, and hence should apply to databases. But in some countries databases are excluded from the scope of the law and in others the law is not clear; practice varies across countries. Similarly, not all countries establish a right of access to information stored in electronic format, and many access to information laws do not make reference to machine readable or open formats. The definition of information in most access to information laws typically refers to all information recorded in any format, which should include databases. However, there is often no explicit reference to a right of access to databases, except for laws such as in Finland and Norway that do expressly permit such access. On the other hand, in Sweden such access is provided but only in printed format, while in the Netherlands and Denmark databases are specifically excluded from the scope of the law. This is a problem predominantly with older access to information laws. In the majority of countries where there is no specific exclusion for databases, access to information and open government data advocates can use the wording of the national access to information law to argue that the right applies to databases.

- *Legal exceptions to openness* – There are a number of ways in which information held by public bodies may – rightfully – not be completely open to information seekers, from a legal perspective. The first is that the information qualifies as a legal exception on grounds such as national security or protection of privacy and is therefore not released to the public, even when someone files an information request. The second is when the public body assesses that the information can be commercialised by being sold to for-profit companies, which can then produce value-added products. The information will therefore be released to members of the public or to private companies only upon payment of a fee. These exemptions are actually necessary to reassure users that the right data are protected; the challenge arises from ensuring that the right criteria are explained to third parties and applied consistently.

- *Complexities of the various national legal frameworks for copyright and related rights as they apply to government digital content databases* – One additional legal area that especially lacks clarity, and that affects public sector information and data "openness", is the question of who owns government data sources and digital content. Many access to information laws presume that public information is to be accessible, and in that sense these laws consider the general public as the legitimate owner of public sector information and data. However, in some countries it is still the case that public bodies assert intellectual property rights such as copyright and database rights over the data they have generated or collected. Even where intellectual property rights are not asserted, public bodies tend to assume that they are the exclusive owners of the data and information, and their economic model sometimes includes selling the information for profit.

- *Compared to technologists in the private sector, national webmasters in the public sector face a daunting array of additional challenges and requirements* – These often are not technology-related, e.g. legal challenges, but they still have an impact on technological matters. In the United States for instance, an online compliance checklist for designers of federal websites identifies about 24 different regulatory regimes with which all public federal websites must comply.[23] These range from privacy and usability to Freedom of Information Act, compliance with the demands of the Paperwork Reduction Act and, separately, the Government Paperwork Elimination Act. Each of these requirements is justified by federal mandate and reflects an assessment informed by the understanding of information technology that was available when it was written. But the cumulative effect of these requirements, taken together, is to place federal web designers in a compliance minefield that makes it hard for them to avoid breaking the rules – while diverting energy from innovation into compliance. These problems are not unique to the United States; they are faced by public websites in many countries.
- *Extent of flexibility in existing regimes* – Updating policies and rules is essential to properly address issues related to putting public sector information on line. A number of recently adopted laws that explicitly address such issues raise a question of interpretation: does an Internet server that contains (machine readable) XML files that can be displayed directly in a web browser and deciphered by humans, but is designed to be used as input into an application count as a "website"? If not, statutory requirements may require government bodies to continue maintaining their own sites. It could be argued that XML pages are not web pages because they cannot be conveniently understood without suitable software to "parse" them and create a human-facing display. Adopting established regulations allowing access to information acts to be operational is important. Furthermore, with access to information acts, the government is expected to promote accessibility to open data for minorities to avoid creating new forms of digital divides, and to increase inclusion. These should include language options for content and access for the disabled, including for the hearing and vision impaired. Inconsistent laws, such as the Official Secrets Act in the United States, if not amended to be brought into line with the requirements of increased transparency and openness by public bodies, can hinder the full-fledged development of OGD initiatives and enforcement of the supporting legislation.

Box 10.9. **Landmark decision in the Netherlands**

In April 2009 the Judicial Division of the Dutch Council of State (*Raad van State*), the highest administrative court in the Netherlands, placed limits on the possibility for public bodies to charge for access to databases they have created, when it ruled that a public authority could not assert database rights over, nor charge for, data collected with public funds as part of its regular activities. The case was taken to the court by Landmark Nederland, a large supplier of land and property search information, which in 2006 put together a national data set of environmental risks such as contaminated land from a range of sources including Dutch council records. These reports were part of a portfolio of products to be sold to homebuyers via estate agency brokers. The City of Amsterdam sought compensation for supplying the data and also wanted to limit its reuse, arguing that a substantial investment had been made in compiling the original data set.

Box 10.9. Landmark decision in the Netherlands (cont.)

The court rejected the appeal lodged by the City of Amsterdam for compensation costs for supplying information that would be sold on for profit. The court ruled that, while the data could be considered to form a database because there had been a substantial investment in its collection, the City of Amsterdam had not borne the risk of this substantial investment, and was therefore not a producer of the database and so could not assert database rights. Consequently, the city was not entitled to attach financial conditions or other limitations on the use of this data by Landmark.

Source: Based on material published on the EPSIPlatform website www.epsiplatform.eu/examples/cases/landmark_nederland_bv_v_amsterdam_city_council, accessed 15 May 2015.

10.3. Key findings and policy conclusions

The public sector is one of the most data-intensive sectors, and is an important actor in the data ecosystem, in two respects: as key user of data and analytics, and as key producer of data. Public sector data can benefit governments (e.g. in terms of public sector productivity and internal costs savings, improved policy development, more effective service delivery, transparency), citizens (e.g. through public participation and engagement, people's empowerment) and businesses (e.g. through product and process innovation).

Objectives related to good governance value – i.e. transparency and government openness – rank among the top motives driving government initiatives to promote open data. This is consistent with the fact that many national agendas on open data emanated as complement to or reinforcement of national transparency agendas. Creating economic value for the private sector also ranks among the top objectives. The objectives of citizen participation and citizen engagement ranked lower than would be expected, given that many governments' open government and service delivery agendas identify open data as a key enabler of strengthened public engagement in serving design, policy making and rulemaking.

The potential of public sector data for the private sector is significant. The OECD market for public sector information (including data) was estimated to be around USD 97 billion in 2008, and could have grown to around USD 111 billion by 2010. Aggregate OECD economic impacts of PSI-related applications and use were estimated to be around USD 500 billion, and there could be close to USD 200 billion of additional gains if barriers to use are removed, skills enhanced and the data infrastructure improved. There is also firm-level evidence that there are significant cross-country benefits from free or marginal cost pricing, with SMEs benefitting most from cheaper data and the switch to marginal cost pricing.

The main barriers to open access to public sector data are not technical but i) policy challenges (e.g. the lack of procedures and standards for dealing with open data in governments), ii) funding challenges (e.g. cost recovery), and iii) organisational and cultural challenges (e.g. ensuring accountability, the quality of data and responsibility in the context of collaboration).

Funding challenges are often seen as a critical challenge at times of budget cuts and financial constraints. Some governments therefore feel the need to clearly articulate a

"business case" and identify funding models. Available evidence suggests, however, that where revenues are collected from the use of public sector data, in most cases they are less than 1% of expenditures, with a maximum of one-fifth of expenditures in a few cases. This suggests that revenue collection models have restricted use without collecting significant revenues. That said, there is a need to establish a clear measure of the potential costs and benefits of opening up public sector data, and to help governments build a better "business case" for open access in the public sector.

A number of countries – including Australia, France, the United Kingdom and the United States – have radically overhauled their open data access systems, and other countries including Norway, the Netherlands and Spain have made access easier and less costly. There are differences in approaches used depending on where countries are positioned in their open data-related activities.[24] Policy strategies include: opening up public sector data that have been difficult to access and reuse; reviewing and amending unnecessary restrictions; reviewing and redefining the public task; facilitating access to third party rights holders' material where rights holders agree. The international dimensions of access to public sector data are also being stressed, both in accessing international data, and in developing international markets for national data.

The following policy options can be recommended based on the discussion in this chapter:

1. Governments should ensure that existing legal and regulatory frameworks enable release of public sector data in open formats and enable non-discriminatory and free-of-charge access and reuse while ensuring the needed level of confidentiality, security and privacy protection.
2. Adopting an overall strategy for public sector data based the strong principles of openness (including machine-readability), copyrights (including standard open data licences such as Creative Commons), and pricing (free or at most marginal cost priced) – and covering issues concerning licences, standards, etc. – should be a priority, to co-ordinate efforts, exploit synergies, facilitate use of linked data, and create a shared view of open data within and across levels of government.
3. Governments should ensure early and timely data release, and the high quality and clarity (i.e. metadata) of published public sector data, as these are all essential conditions to enable reuse and value creation.
4. Recognising that the public sector holds a vast amount of data and information that may be of interest to the public, governments should improve their knowledge of the needs of the community of users and their capacity to consult with them to identify which data to prioritise for release as open data.
5. Governments should nurture the development of the data ecosystem and promote a culture of collaboration among the key actors to increase the value created from public sector data. A wide range of public sector data reuse by a wide range of actors is a key condition for economic and social value creation, and necessary to stimulate creativity and innovation.
6. Governments should increase open data literacy, within both the public sector and society, to promote reuse and thus unlock the value of open data.
7. Governments should promote coherence among open data frameworks, many of which relate to access, linkage and reuse. In this respect, the merging of existing

OECD Council Recommendations that aim to promote better access to and use of data could be considered to stimulate a data-driven public sector. These recommendations include the OECD (2008) *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* of 30 April 2008, the OECD (2006) *Recommendation of the Council concerning Access to Research Data from Public Funding* of 14 December 2006 – both currently under review – and the OECD (2014d) *Recommendation of the Council on Digital Government Strategies*.

Annex – Principles of the OECD (2008) Council Recommendation on PSI

- **Openness**. Maximising the availability of public sector information for use and reuse based upon presumption of openness as the default rule to facilitate access and reuse. Developing a regime of access principles or assuming openness in public sector information as a default rule wherever possible no matter what the model of funding is for the development and maintenance of the information. Defining grounds of refusal or limitations, such as for protection of national security interests, personal privacy, preservation of private interests for example where protected by copyright, or the application of national access legislation and rules.
- **Access and transparent conditions for reuse**. Encouraging broad non-discriminatory competitive access and conditions for reuse of public sector information, eliminating exclusive arrangements and removing unnecessary restrictions on the ways in which it can be accessed, used, reused, combined or shared, so that in principle all accessible information would be open to reuse by all. Improving access to information over the Internet and in electronic form. Making available and developing automated on-line licensing systems covering reuse in those cases where licensing is applied, taking into account the copyright principle below.
- **Asset lists**. Strengthening awareness of what public sector information is available for access and reuse. This could take the form of information asset lists and inventories, preferably published on-line, as well as clear presentation of conditions to access and reuse at access points.
- **Quality**. Ensuring methodical data collection and curation practices to enhance quality and reliability including through co-operation of various government bodies involved in the creation, collection, processing, storing and distribution of public sector information.
- **Integrity**. Maximising the integrity and availability of information through the use of best practices in information management. Developing and implementing appropriate safeguards to protect information from unauthorised modification or from intentional or unintentional denial of authorised access to information.
- **New technologies and long-term preservation**. Improving interoperable archiving, search and retrieval technologies and related research including research on improving access and availability of public sector information in multiple languages, and ensuring development of the necessary related skills. Addressing technological obsolescence and challenges of long-term preservation and access. Finding new ways for the digitisation of existing public sector information and content, the development of born-digital public sector information products and data, and the implementation of cultural digitisation projects (public broadcasters, digital libraries, museums, etc.) where market mechanisms do not foster effective digitisation.
- **Copyright**. Intellectual property rights should be respected. There is a wide range of ways to deal with copyrights on public sector information, ranging from governments or private entities holding copyrights, to public sector information being copyright-free. Exercising copyright in ways that facilitate reuse (including

waiving copyright and creating mechanisms that facilitate waiving of copyright where copyright owners are willing and able to do so, and developing mechanisms to deal with orphan works), and where copyright holders are in agreement, developing simple mechanisms to encourage wider access and use (including simple and effective licensing arrangements), and encouraging institutions and government agencies that fund works from outside sources to find ways to make these works widely accessible to the public.

- **Competition**. Ensuring that pricing strategies take into account considerations of unfair competition in situations where both public and business users provide value-added services. Pursuing competitive neutrality, equality and timeliness of access where there is potential for cross-subsidisation from other government monopoly activities or reduced charges on government activities. Requiring public bodies to treat their own downstream/value-added activities on the same basis as their competitors for comparable purposes, including pricing. Particular attention should be paid to single sources of information resources. Promoting non-exclusive arrangements for disseminating information so that public sector information is open to all possible users and reusers on non-exclusive terms.
- **Redress mechanisms**. Providing appropriate transparent complaints and appeals processes.
- **Public private partnerships**. Facilitating public-private partnerships where appropriate and feasible in making public sector information available, for example by finding creative ways to finance the costs of digitisation, while increasing access and reuse rights of third parties.
- **International access and use**. Seeking greater consistency in access regimes and administration to facilitate cross-border use and implementing other measures to improve cross-border interoperability, including in situations where there have been restrictions on non-public users. Supporting international co-operation and co-ordination for commercial reuse and non-commercial use. Avoiding fragmentation and promote greater interoperability and facilitate sharing and comparisons of national and international data sets. Striving for interoperability and compatible and widely used common formats.
- **Best practices**. Encouraging the wide sharing of best practices and exchange of information on enhanced implementation, educating users and reusers, building institutional capacity and practical measures for promoting reuse, cost and pricing models, copyright handling, monitoring performance and compliance, and their wider impacts on innovation, entrepreneurship, economic growth and social effects.

Notes

1 The survey was undertaken by the OECD Directorate for Public Governance and Territorial Development (GOV). A complementary survey by the OECD Directorate for Science, Technology and Innovation was undertaken in parallel but focused on the larger concept of PSI with the review of the OECD Council.

2 It is necessary to exercise caution when interpreting these results, as the methodologies used for these estimates are unknown.

3 Full information on the data included can be found at: www.ic.nhs.uk/services/national-clinical-audit-support-programme-ncasp/diabetes, accessed 15 May 2015.

4 UN Globalpulse introduced the concept of "data philanthropy", whereby the private sector shares data to support more timely and targeted policy action, and to heighten public interest in shared data. In this context, two ideas are debated: i) the "data commons", where some data are shared publicly after adequate anonymisation and aggregation; and ii) the "digital smoke signals", where companies share the results of sensitive data with government but not the data themselves.

5 Askitas and Zimmermann (2009), for example, analyse the predictive power of keywords such as Arbeitsamt OR Arbeitsagentur ("unemployment office or agency") to forecast unemployment in Germany. The authors find that forecasting based on these keywords indicated changes in trends much earlier than official statistics. Similar conclusions have been drawn by D'Amuri and Marcucci (2010) for the United States and by Suhoy (2010) for Israel.

6 See www1.unece.org/stat/platform/display/bigdata/Big+Data+Project, accessed 15 May 2015.

7 Open Government Declaration, "Open Government Partnership" (September 2011), www.opengovernmentpartnership.org/sites/www.opengovernmentpartnership.org/files/page_files/OGP_Declaration.pdf, signed by the United States and seven other countries in September 2011.

8 In contrast to Web 1.0 applications, which were conceived for the passive delivery of content to a mass audience broadcast from 'one-to-many', Web 2.0 applications allow users to participate directly in the creation, refinement and distribution of shared content (user-created content, UCC) (see OECD, 2007).

9 For example, the Department of Health and Human Services in the United States has pushed for the "smart disclosure" of data on flights operated by national airlines, to enable people to make informed choices on the airline company selection.

10 Also identified were the main providers among federal departments, agencies and offices of data used by the 500 companies studied. The main providers appear to be Department of Commerce and Department of Health and Human Services, followed (distantly) by the Securities and Exchange Commission, the Department of Labor, the Department of Energy, the Department of Education, and the Environmental Protection Agency.

11 See http://europa.eu/about-eu/facts-figures/economy/index_en.htm, and http://en.wikipedia.org/wiki/Economy_of_the_European_Union, accessed 10 May 2013.

12 Spatial information is around one-half of all PSI according to PIRA, 2000; MEPSIR, 2006; and Proyecto Aporta, 2011.

13 Note that these values differ somewhat from those estimated in previous work (Vickery, 2011, 2012), due to the use of more recent macroeconomic data and the choice of exchange rates to convert national estimates to USD and EUR.

14 Countries generally have the stated aim of being able to provide standardised and appropriately comprehensive metadata with all data sets, but most central portals fall short of this aim. This is due to the reliance on making available existing data sets that may not have extensive, or any, associated metadata.

15 A field in a database that contains personal data such as the e-mail addresses of private individuals can be removed before the remainder of the information is released, in order to protect personal privacy while respecting the right of access to information.

16 Most of these countries provide open data via participation and collaboration platforms – United States: www.data.gov; United Kingdom: www.data.gov.uk; Australia: www.data.gov.au; France: www.data.gouv.fr, accessed 30 July 2012.

17 In order to establish a framework for fair, proportionate and non-discriminatory conditions for the reuse of information held by public sector bodies in the European Union, the European Commission adopted Directive 2003/98/EC, which states in Article 1 that its main objective is to establish "a minimum set of rules governing the re-use and the practical means of facilitating re-use of existing documents held by public sector bodies of the Member States". This objective should be placed in the context of the wider goal of facilitating access to knowledge for citizens, and business promoting the emergence of Community-wide (data-driven) services as an important part of the internal market: http://ec.europa.eu/information_society/policy/psi/index_en.htm, accessed 14 May 2015.

18 The German Law on the reuse of information for public bodies ("Informationsweiterverwendungsgesetz"), implemented in December 2006, reflects the aims and goals of the EU PSI Directive. However, it does not include elements to proactively provide government data to the public, nor does it create the right of access to government information; application of the law assumes such a right is already in place. As a result, the decision as to whether official information may be reused and the details of that use are subject to the discretion of the public authority

concerned (Schellong and Stepanets, 2011). source Unchartered Waters – The State of Open Data in Europe, Business Solutions Technology Outsourcing, 2011).

19 See www.aporta.es/web/guest/guia_reutilizacion, accessed 14 May 2014.

20 See www.gfii.asso.fr/article.php3?id_article=3278, accessed 14 May 2014.

21 See www.digitaliser.dk/resources/559456, accessed 14 May 2014.

22 A government database of public bodies' payments and transactions, at www.rgs.mef.gov.it/ENGLISH-VE/SIOPE1/, accessed 14 May 2014.

23 Web Content Managers Advisory Council, Requirements Checklist for Government Web Managers, www.usa.gov/webcontent/reqs_bestpractices/reqs_checklist.shtml, accessed 2 December 2008.

24 The exchange of experiences and best practices is crucial for the development of more ambitious and innovative action plans related to open data. The International Open Data Working Group, currently chaired by Canada and working in the context of the Open Government Partnership, offers a platform for governments to share successes, failures and new ideas.

References

ACIL Tasman (2009), "Spatial information in the New Zealand economy: Realising productivity gains", prepared for Land Information New Zealand; Department of Conservation; Ministry of Economic Development, www.geospatial.govt.nz/productivityreport, accessed 15 May 2015.

ACIL Tasman (2008), "The value of spatial information: The impact of modern spatial information technologies on the Australian economy", report prepared for the CRC for Spatial Information and ANZLIC, Australia, the Spatial Information Council, www.anzlic.org.au/Publications/Industy/251.aspx, accessed 15 May 2015.

Askitas N. and KN. Zimmermann (2010), "Google econometrics and unemployment forecasting", Technical report, SSRN 899, 2010, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341, accessed 13 May 2015.

Castelein, W.T., A. Bregt and Y. Pluijmers (2010), "The economic value of the Dutch geo-information sector", *International Journal of Spatial Data Infrastructures Research*, Vol. 5, pp. 58-76.

Carrière-Swallow, Y. and F. Labbé (2010), "Nowcasting with Google Trends in an Emerging Market", *Central Bank of Chile Working Papers*, No. 588, July, www.bcentral.cl/estudios/documentos-trabajo/pdf/dtbc588.pdf, accessed 13 May 2015.

Cebr (2012), "Data equity: Unlocking the value of big data", Report for SAS, April, www.sas.com/offices/europe/uk/downloads/data-equity-cebr.pdf, accessed 13 May 2015.

Coote, A. and A. Smart (2010), "The value of geospatial information to local public service delivery in England and Wales", Local Government Association, www.lga.gov.uk/GIresearch, accessed 15 May 2015.

Craglia, M., L. Pavanello and R. S. Smith (2010), "The Use of Spatial Data for the Preparation of Environmental Reports in Europe", European Commission Joint Research Centre Institute for Environment and Sustainability, Ispra, Italy, available at: http://ies.jrc.ec.europa.eu/uploads/SDI/publications/JRC_technical%20report_2009%20EIA-SEA%20survey.pdf, accessed 13 May 2015.

D'Amuri, F. and J. Marcucci (2010), "Google it! Forecasting the US unemployment rate with a Google job search index", SSRN, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1594132, accessed 14 May 2015.

Danish Enterprise and Construction Authority (2010), "The value of Danish address data", 7 July.

Davis, T. (2010), "Open data, democracy and public sector reform: A look at open government data use at data.gov.uk", Thesis, University of Oxford, www.opendataimpacts.net/report/, accessed 13 May 2015.

European Statistical System Committee [ESSC] (2013), *Scheveningen Memorandum on "Big Data and Official Statistics",* September, available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version_0.pdf, accessed 13 May 2015.

Farina, C. et al. (2008), "Achieving the potential: The future of federal e-rulemaking", Committee on the Status and Future of Federal e-Rulemaking (US), http://scholarship.law.cornell.edu/facpub/1237/, accessed 13 May 2015.

Federal Department of the Interior (Switzerland), (2013), "On the economic impact of open government data", English summary available at: www.opendata.admin.ch/en/about, accessed 14 May 2015.

Fornefeld, M. (2011), "INSPIRE & open data: Activator for the European PSI market?", presentation at Open Data: Apps for Everyone? Opportunities and Challenges in the Reuse of Public Sector Information", Berlin, 18 February.

Howard, A. (2012a), "Predictive data analytics is saving lives and taxpayer dollars in New York City", *O'Reilly Radar*, 26 June, http://strata.oreilly.com/2012/06/predictive-data-analytics-big-data-nyc.html, accessed 8 May 2015.

Howard, A. (2012b), "What is Smart Disclosure?", *O'Reilly Radar*, 1 April, http://radar.oreilly.com/2012/04/what-is-smart-disclosure.html, accessed 13 May 2015.

Kalampokis, E., E.Tambouris and K. Tarabanis (2012), "A classification scheme for open government data: Towards linking decentralised data", *International Journal of Web Engineering and Technology,* Volume 6 Issue 3, June 2011, pp. 266-285, http://dl.acm.org/citation.cfm?id=1999591.

Koski, H. (2011), "Does marginal cost pricing of public sector information spur firm growth?", *ETLA Discussion Papers*, No. 1260, www.etla.fi/en/publications/dp1260-en, accessed 15 May 2015.

MGI (McKinsey Global Institute) (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company, June, www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx , accessed 15 May 2015.

Measuring European Public Sector Information Resources [MEPSIR] (2006), "Final report of study on exploitation of public sector information – benchmarking of EU framework conditions", Executive summary, final report Part 1 and Part 2, http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1197 http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1198 http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1199, accessed 13 May 2015.

MICUS (2010), "European legislation as a driver for German geobusiness", MICUS Management Consulting, Dusseldorf, Germany, www.micus.de/51a_GeoBusiness_en.html, accessed 15 May 2015.

MICUS (Fornefeld, M., G. Boele-Keimer, S. Recher and M. Fanning) (2009), "Assessment of the reuse of public sector information (PSI) in the geographical

information, meteorological information and legal information sectors", MICUS Management Consulting, Dusseldorf, Germany.

Millard, J. (2012), "Rethinking e-participation: Smash down the silos and move to open participation", white paper prepared for the E-Participation Summit in Stockholm, 9-11 May, on behalf of the Swedish Association for Local Authorities and Regions, http://skl.se/download/18.9f425ef147b396d4676d20d/1408369822233/Whitepaper%26Slideshow_Jeremy_Millard.pdf, accessed 15 May 2015.

OAIC (2013), "Open public sector information: From principles to practice – Report on agency implementation of the Principles on Open Public Sector Information", Office of the Australia Information Commissioner, February.

OECD (2015), "Assessing government initiatives on public sector information: A review of the OECD Council Recommendation", OECD, forthcoming.

OECD (2014a), *International Forum on Open Government – Agenda*, OECD Publishing, Paris, 30 September, www.oecd.org/mena/governance/FINAL%20Agenda%20Forum%20Open%20Gov%20ENG.pdf, accessed 15 May 2015.

OECD (2014b), "OECD.StatExtracts", *National Accounts*, OECD Publishing, Paris.

OECD (2014c), Spain: From administrative reform to continuous improvement, *OECD Public Governance Reviews*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264210592-en.

OECD (2014d), Recommendation of the Council on Digital Government Strategies, 15 July 2014, C(2014)88, OECD Publishing, Paris, www.oecd.org/gov/public-innovation/recommendation-on-digital-government-strategies.htm.

OECD (2013), "Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by 'Big Data'", *OECD Digital Economy Papers*, No. 222, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k47zw3fcp43-en.

OECD (2008), Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information, C(2008)36, 30 April 2008, OECD, www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2006), Recommendation of the Council concerning Access to Research Data from Public Funding, 14 December 2006, C(2006)184, OECD Publishing, Paris.

OMB (2004), "Expanding e-government: Partnering for a results oriented government", US Office of Management and Budget, Executive Office of the President, www.whitehouse.gov/omb/budintegration/expanding_egov12-2004.pdf, accessed 14 May 2015.

Pira (2000), *Commercial Exploitation of Europe's Public Sector Information*, *Executive Summary*, Pira International Ltd, University of East Anglia and KnowledgeView Ltd, and *Final Report*, Pira International, European Commission, Directorate General for the Information Society.

Pollock, R. (2010), "Welfare gains from opening up public sector information in the UK", University of Cambridge, 15 September, http://rufuspollock.org/economics/papers/psi_openness_gains.pdf, accessed 15 May 2015.

Proyecto Aporta (2011), "Characterization Study of the Infomediary Sector", prepared by the Ministry of Territorial Policy and Public Administration, the State Secretariat of Telecommunications and Information Society, of the National Observatory of Telecommunications, of the Information Society (ONTSI), and of the Ministry of Industry, Tourism and Trade, Madrid, www.aporta.es/web/guest/estudioRISP2011; updated (2012) as "Characterization Study of the Infomediary Sector 2012 Edition", http://datos.gob.es/datos/?q=node/2148, accessed 10 May 2015.

Regulations.gov (n.d.), www.regulations.gov/#!home, accessed 9 May 2015.

Reimsbach-Kounatze, C. (2015), "The proliferation of data and implications for official statistics and statistical agencies: A preliminary analysis", *OECD Digital Economy Working Papers*, http://dx.doi.org/10.1787/5js7t9wqzvg8-en.

Robinson, D. and H. Yu (2012), "The New Ambiguity of Open Government", *UCLA Law Review, Discourse*.

Schellong, A. and E. Stepanets (2015), "Unchartered waters: The state of open data in Europe", *CSC Public Sector Study Series*, http://assets1.csc.com/de/downloads/CSC_policy_paper_series_01_2011_unchartered_waters_state_of_open_data_europe_English_2.pdf, accessed 9 May 2015.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of government in a digital age", Computer and Communications Industry Association, October, www.ccianet.org/CCIA/files/ccLibraryFiles/Filename/000000000086/govtcomp_report.pdf, accessed 10 October 2013.

Suhoy, T. (2009), "Query indices and a 2008 downturn: Israeli data", Technical Report, Bank of Israel, 2009, www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf, accessed 13 May 2015.

Surowiecki, J. (2011), "A Billion Prices Now", *The New Yorker*, 30 May, p. 28, www.newyorker.com/magazine/2011/05/30/a-billion-prices-now, accessed 13 May 2015.

Ubaldi, B. (2013), "Open government data: Towards empirical analysis of open government data initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

United Nation [UN] Global Pulse (2012), "Big Data for Development: Opportunities & Challenges", Global Pulse White Paper, May, available at: www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf, accessed 13 May 2015.

Verdier, H. (2014), "Open data et démocratie : la réponse d'Henri Verdier, « monsieur Data » du gouvernement", 11 August, http://rue89.nouvelobs.com/2014/08/11/open-data-democratie-reponse-dhenri-verdier-m-data-gouvernement-254156, accessed 14 May 2015.

Vickery, G. (2012), "Review of recent studies on PSI reuse and related market developments", Update, Statistivcs Sweden, www.scb.se/statistik/_publikationer/NR9999_2012A01_BR_X76BR1201.pdf, accessed 10 May 2015.

Vickery, G. (2011), "Review of recent studies on PSI reuse and related market developments", European Commission,

http://ec.europa.eu/information_society/policy/psi/facilitating_reuse/economic_analysis/index_en.htm, accessed 10 May 2015.

WITSA (2009), "Digital Planet 2009: Report tables", World Information Technology and Services Alliance, based on research conducted by Global Insight Inc., Vienna, Virginia.

Further reading

De Vries, M. (2012), "Reuse of public sector information – Catalogue and highlights of studies, cases and key figures on economic effects of changing policies", Report for the Danish Ministry for Housing, Urban and Rural Affairs, 11 August.

European Union (2013), Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the reuse of public sector information. For an overview see http://ec.europa.eu/digital-agenda/en/legal-rules, accessed 15 May 2015.

European Union (2003), Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the reuse of public sector information.

Etalab (2013), "Etalab a été rattachée directement au Secrétaire général pour la modernisation de l'action publique", www.etalab.gouv.fr, accessed 15 May 2015

Etalab (2011), "Création de la mission Etalab, chargée de la mise en ligne de data.gouv.fr", www.gouvernement.fr/gouvernement/creation-de-la-mission-etalab-chargee-de-la-mise-en-ligne-de-datagouvfr, accessed 15 May 2015.

Ministry of Government Administration, Reform and Church Affairs (2013), "Digital Agenda for Norway", white paper, 22 March, www.regjeringen.no/pages/38354256/PDFS/STM201220130023000EN_PDFS.pdf, accessed 15 May 2015.

OAIC (2012), "Information publication scheme and public sector information: Survey of Australian government agencies; Data pack report for Part B: Management and publication of public sector information", Office of the Australian Information Commissioner, Orima Research, Australia.

OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris.

OMB (2009), "Open Government Directive", US Office of Management and Budget, Executive Office of the President, 8 December, www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf, accessed 15 May 2015.

Pollock, R. (2011), "Funding options for trading funds and other PSI holders", accessed 1 March 2011, http://rufuspollock.org/economics/papers/psi-funding-options/ via http://rufuspollock.org/economics/, accessed 15 May 2015.

Pollock, R. (2009), "The economics of public sector information", University of Cambridge, 2 December 2008, subsequently published as *Cambridge Working Papers in Economics,* 0920, Faculty of Economics, University of Cambridge, www.rufuspollock.org/economics, accessed 15 May 2015.

Pollock, R., D. Newbery and L. Bently (2008), "Models of public sector information provision via trading funds", Department for Business, Enterprise and Regulatory Reform, United Kingdom.

Power of Information Taskforce (2009), "Final report", 13 April 2010, http://webarchive.nationalarchives.gov.uk/20100413152047/http://poit.cabinetoffice.gov.uk/poit/2009/02/summary-final/, accessed 15 May 2015.

SerdaLAB (2012), "Information électronique professionnelle: Marché et tendances en 2011-2012", SerdaLAB, Paris.

SerdaLAB (2010), "L'information électronique professionnelle en France: Le marché et les tendances en 2009-2010", SerdaLAB, Paris.

The GovLab (2014), *Open Data 500*, The Governance Lab @ NYU, http://thegovlab.org/tag/open-data-500/, accessed 10 May 2015.

The National Archives (2011),"The United Kingdom Report on the Reuse of Public Sector Information 2010: Unlocking PSI potential", 1 April, www.nationalarchives.gov.uk/documents/psi-report.pdf, accessed 15 May 2015.

UK Cabinet Office (2012), "Open data white paper: Unleashing the potential", June.

Zangenberg & Company (2011), *Kvantificering af værdien af åbne offentlige data,* [Quantifying the value of open government data], Report prepared for the Danish National IT and Telecom Agency, http://digitaliser.dk/resource/1021067/artefact/Kvantificering+af+den+erhvervsm%C3%A6ssige+v%C3%A6rdi+af+%C3%A5bne+offentlige+data+-+Zangenberg2011.pdf, accessed 15 May 2015.

Glossary

Knowledge is understood as *information and experience internalised or assimilated* through a process, commonly referred to as "learning". It provides the "learner" with the capacity to make effective decisions autonomously. Knowledge can be explicit, in which case it can be cost-effectively externalised to be communicated and embedded in tangible products, including books, standard procedures and intangible products such as patents, design and software. But it can also be tacit, based on an "amalgam of information and experience", which is too costly to codify and thus to externalise.

Information is often seen as the *meaning* resulting from the interpretation of facts as conveyed through *data* or other sources such as words. This meaning is reflected in the structure or organisation of the underlying source, including its hidden relationships and patterns of correlations, which can be revealed through *data analytics*. Information is therefore always context-dependent: it depends on the capacity to extract meaning from the information source; this capacity depending on available data analytic techniques and technologies as well as the skills and (pre-)knowledge of the data analyst.

Data are understood as the *representation of facts* stored or transmitted as qualified or quantified symbols. Data have no inherent meaning; however, they can be domain-specific. In contrast to knowledge and information, data are assumed to have an "objective existence", and they can be measured, namely in bits and bytes (see Table below). Data are typically gained from information when that information is *encoded* so it can be stored or communicated. Data can also be the result of *datafication*, a portmanteau for "data" and "quantification", where a phenomenon or object is transformed into quantified symbols. Datafication should not be confused with *digitisation*, which refers to the process of encoding information into *binary digits* (i.e. bits) so it can be processed by computers. Data that have not been digitised cannot be processed by computers.

Big data initially referred to data for which the i) *volume* became an issue in terms of data management and processing. Further definitions highlighted other important characteristics of "big data", such as ii) *velocity,* or the speed at which data are generated, accessed, processed and analysed (referring to real-time data), and iii) *variety* (referring to *unstructured* data and the capacity to link diverse data sets). These three properties – volume, velocity and variety – are therefore often considered to be the three main characteristics, and are commonly referred to as the three Vs, of big data. There is a major limitation with definitions based on the 3Vs, however: they are in continuous flux, as they describe technical properties that depend on the evolving state of the art in data storage and processing. Furthermore, these definitions misleadingly suggest that data are the main source of value. While it is true in the case of volume, what is behind variety and velocity is primarily *data analytics* – that is, the capacity to analyse unstructured diverse data in (close to) real time. Furthermore the term "big data" does not suggest how the data are used what type of innovation they can enable, or a how they relate to other concepts such as (e.g.) *open data*, *linked data*, and *data mashups*.

Units for measuring the volume of data

Unit	Size	What it means
Bit (B)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers uses to store and process data.
Byte (B)	8 bits	Enough information to create a number or an English letter in computer code. It is the basic unit of computing.
Kilobyte (KB)	1 000 B	From "thousand" in Greek. One page of typed text is 2 KB.
Megabyte (MB)	1 000 KB	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4 MB.
Gigabyte (GB)	1 000 MB	From "giant" in Greek. A two-hour film can be compressed into 1-2GB.
Terabyte (TB)	1 000 GB	From "monster" in Greek. All the catalogued books in the US Library of Congress total around 15 TB.
Petabyte (PB)	1 000 TB	All letters delivered by America's postal service in 2011 will amount to around 5 PB; Google processes around 1 PB every hour.
Exabyte (EB)	1 000 PB	Equivalent to 10 billion copies of *The Economist.*
Zettabyte (ZB)	1 000 EB	The total amount of information in existence in 2011 was around 1.2 ZB.
Yottabyte (YB)	1 000 ZB	Currently too big to imagine.

Note: The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Source: Adopted from *The Economist (2010), "Data, data everywhere", The Economist, 25 February,* www.economist.com/node/15557443.

Structured data are data based on a predefined *data model* (i.e. an abstract representation of "real world" objects and phenomenon). Such models can be explicit, as in the case of a structured query language (SQL) database, where the data model is reflected in the structure of the database's tables. The data model can also be implicit, as in the case of *semi-structured data* (e.g. structured web content), where the underlying model can be made explicit at relatively low cost. In contrast, **unstructured data** are data that have no predefined data model and where such a model cannot be cost-effectively extracted. Typical examples include text-heavy data sets such as text documents and e-mails, as well as multimedia content such as videos, images and audio streams. The difference between structured, semi-structured, and unstructured data is becoming less important since with rising computing capacities, *data analytics* are increasingly able to automatically extract some structures embedded in unstructured data, including multimedia content.

Linked data typically refers to structured data that are published so that they can be interlinked. Data linkage is a means to contextualise data and thus enable the extraction of further information, which is greater than the sum of the information from the isolated **data silos**. The concept of linked data is closely related to the concept of open data, for which the full benefits can only be achieved if the isolated open data sets can be interlinked. Open standards play an important role in an interlinked data ecosystem.

Metadata are data about entities, including (**primary) data**. Metadata provide the necessary context without which the primary data cannot be accessed, linked, or fully understood. Metadata can be i) descriptive (based on attributes used to search and find an entity), ii) structural (describing the structure and organisation of an entity such as databases), and iii) administrative (providing information to help manage a resource). The concept of metadata is closely related to the concept of linked data, since metadata and primary data are by definition linked.

Personal data are defined by the OECD *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* as "any information relating to an identified or identifiable individual (data subject)". Any data that are not related to an identified or identifiable individual are therefore "non-personal" data. However, *data analytics* has made it easier to relate seemingly non-personal data to an identified or identifiable individual, thus blurring the boundaries between non-personal and personal data (see Chapter 5). It should be noted that the definition of personal data applied here does not distinguish between data (as inherently meaningless representation of facts) and information (as the *meaning* resulting from the interpretation of data). In other words, personal data and personal information are used as synonyms in this report.

Data can be **volunteered** when they are explicitly shared (by a data subject). Examples include creating a social network profile and entering credit card information for online purchases. They can be **observed** when it is captured by recording activities. In contrast to volunteered data where the data subject is actively and purposefully sharing its data, the role of the observed data subject is passive. Examples of observed data include location data of cellular mobile phones, and web usage behaviour. And finally, information can be **inferred** as the result of *data analytics*. Examples include credit scores calculated based on an individual's financial history. It is interesting to note that personal information can be "inferred" from several pieces of seemingly "anonymous" or "non-personal" data.

Public sector (government) data, in respect to the OECD *Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* (PSI), are data generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the government or public institutions (see Chapter 10). They are: i) dynamic and continually generated, ii) often directly produced by the public sector, or iii) associated with the functioning of the public sector (e.g. meteorological data, geo-spatial data, business statistics), and iv) often readily useable in commercial applications with relatively little transformation, as well as being the basis of extensive elaboration. Public sector data are a subset of PSI, which includes not only data but also *digital content,* such as text documents and multimedia files. The terms "public sector data" and "government data" are used as synonyms. The often used term "open government data" refers to public sector data made available as *open data*.

Open data does not describe a specific type of data. The key characteristic is the attribute "open", which specifies how access to data is *managed*, namely on *non-discriminatory terms* or "access on equal terms" as stated in the OECD *Recommendation of the Council on Principles and Guidelines for Access to Research Data from Public Funding*. In other words, data become "open" when access is not limited based on users' identity or intended use of the data (see Chapter 4). "Openness" should not be understood as a binary attribute but rather as a *continuum,* ranging from i) *closed* (with access only by e.g. the data controller or data subject), to ii) *commons* with possible restriction to a community (e.g. of researchers), to iii) *(unlimited) access granted to the public* as the highest degree of openness. Three key factors affect the degree of openness:

- technological design (including e.g. availability, machine readability and interoperability)
- intellectual property rights (IPRs) (including copyright as well as other IPRs applicable to databases and trade secrets)

- pricing, with marginal cost pricing being recommended by the OECD (2006) Council Recommendation on Access to Research Data from Public Funding and the OECD *Recommendation of the Council on Enhanced Access and More Effective Use of PSI*.

Data analytics refers to the set of techniques and tools used to extract information from data by revealing the context in which the data are embedded, their organisation and their structure. In the case of visual analytics the emphasis lies on data visualisation including (interactive) data exploration. Data analytics reveals the signal from the noise and with that the data's manifold hidden relations (patterns) including correlations, and interactions between facts, entities, and concepts. A number of terms are used (in this volume as synonyms) to refer to data analytics, some of which may include aspects that go beyond data analysis:

- **Data (text) mining** and **knowledge discovery** typically refer to data analysis but include aspects such as data pre-processing (cleaning), as well as model and inference considerations.
- **Profiling** is often used to describe the construction of profiles and the classification of entities in specific profiles.
- **Business intelligence**, a term that refers to the analysis of business-related data as often stored in databases (data warehouses) and mainly used for business reporting and monitoring purposes.
- **Machine** or **statistical learning** is a subfield in computer science, and more specifically artificial intelligence (AI), concerned with the design, development and use of data analytic algorithms that allow computers to "learn" – that is, to improve performance with every data set analysed.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

The OECD is a unique forum where governments work together to address the economic, social and environmental challenges of globalisation. The OECD is also at the forefront of efforts to understand and to help governments respond to new developments and concerns, such as corporate governance, the information economy and the challenges of an ageing population. The Organisation provides a setting where governments can compare policy experiences, seek answers to common problems, identify good practice and work to co-ordinate domestic and international policies.

The OECD member countries are: Australia, Austria, Belgium, Canada, Chile, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Japan, Korea, Luxembourg, Mexico, the Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom and the United States. The European Union takes part in the work of the OECD.

OECD Publishing disseminates widely the results of the Organisation's statistics gathering and research on economic, social and environmental issues, as well as the conventions, guidelines and standards agreed by its members.

OECD PUBLISHING, 2, rue André-Pascal, 75775 PARIS CEDEX 16
(93 2015 02 1 P) ISBN 978-92-64-22934-1 – 2015

Data-Driven Innovation

BIG DATA FOR GROWTH AND WELL-BEING

Contents

Chapter 1. The phenomenon of data-driven innovation

Chapter 2. Mapping the global data ecosystem and its points of control

Chapter 3. How data now drive innovation

Chapter 4. Drawing value from data as an infrastructure

Chapter 5. Building trust for data-driven innovation

Chapter 6. Skills and employment in a data-driven economy

Chapter 7. Promoting data-driven scientific research

Chapter 8. The evolution of health care in a data-rich environment

Chapter 9. Cities as hubs for data-driven innovation

Chapter 10. Governments leading by example with public sector data

Consult this publication on line at ***http://dx.doi.org/10.1787/9789264229358-en.***

This work is published on the OECD iLibrary, which gathers all OECD books, periodicals and statistical databases.
Visit ***www.oecd-ilibrary.org*** for more information.

OECD*publishing*
www.oecd.org/publishing

ISBN 978-92-64-22934-1
93 2015 02 1 P

